

Misura di grandezze elettriche in continua

francesco.fuso@unipi.it

(Dated: version 8b - FF, 11 novembre 2019)

Questa nota ha lo scopo principale di introdurre i principi di operazione, le caratteristiche funzionali e alcuni dettagli operativi degli strumenti di misura di tensioni e correnti continue. Queste misure sono spesso eseguite usando degli strumenti multifunzionali detti multimetri, o tester. Essi consentono anche una misura della resistenza elettrica, a cui anche si fa cenno in questa nota. Gli aspetti cruciali e non banali degli argomenti considerati sono: il concetto di misura analogica o digitale, la perturbazione che gli strumenti producono sul sistema sotto analisi, modellata attraverso la loro resistenza interna, e la determinazione delle incertezze di misura.

I. INTRODUZIONE

La misura di grandezze elettriche *in continua*, cioè in condizioni costanti o stazionarie, è una prassi comunissima nell'analisi di circuiti elettrici o elettronici e, più in generale, per la pratica sperimentale: al giorno d'oggi pressoché qualsiasi grandezza fisica viene misurata attraverso sensori o trasduttori che forniscono in uscita grandezze elettriche (dette, qui e in previsione di situazioni non necessariamente stazionarie, anche *segnali* elettrici). In questa nota ci occupiamo in particolare di differenze di potenziale, o *tensioni*, ΔV , e di intensità di corrente I , con qualche cenno alla misura di resistenza, R .

Fare il corretto uso di uno strumento multifunzione in grado di compiere tali misure (*multimetro* o *tester*) non è mai un'operazione ovvia o scontata. Alcune considerazioni generali sull'impiego del tester sono elencate qui di seguito.

- A causa della loro multifunzionalità e della possibilità di scegliere diverse portate (fondo scala), i tester devono essere impiegati *sapendo* cosa si deve misurare e, se possibile, avendo un'idea dell'ordine della grandezza che si vuole misurare. Essi vanno dunque usati con la testa, e non con le mani (o con i piedi), pena misure erronee e possibili danneggiamenti.
- È bene ricordare sempre che l'uso dello strumento "perturba" il circuito sotto analisi, ovvero il sistema che si sta misurando. È indispensabile capire bene come può essere modellata la perturbazione e, sulla base di questo modello, tenere in conto (quantitativamente) dei suoi effetti, o verificare in quali condizioni essa può eventualmente essere trascurata.
- L'incertezza, cioè la barra di errore da attribuire alla misura, deve essere valutata con grande attenzione, tenendo conto nella stima, in generale, di tutti gli effetti coinvolti.

II. MISURE DI TENSIONI E CORRENTI (E RESISTENZE)

In genere un tester può misurare tensioni e correnti, sia in continua (*DC* per direct current) che in alternata (*AC* per alternate current), oltre a resistenze e, spesso, anche numerose altre grandezze. Affinché esso possa essere impiegato per la misura di una determinata grandezza occorrono alcuni passaggi preliminari. In particolare si deve:

1. configurare lo strumento per la misura richiesta *prima* della misura (se possibile, prima di fare qualsiasi collegamento elettrico). Per esempio, il tester digitale ha una manopola (collegata a un commutatore) che permette di selezionare la grandezza misurata e anche la sua scala (o portata). Inoltre anche la scelta delle boccole a cui vanno collegati i fili che dallo strumento vanno al circuito, o viceversa, dipende dall'esigenza di misurare tensioni o correnti. Per il tester analogico in uso in laboratorio, la configurazione si fa solo tramite la scelta delle boccole, che quindi diventa cruciale.
2. Scegliere la portata in maniera opportuna. Impiegare la portata "sbagliata" ha almeno due conseguenze negative: (i) se il fondo scala è piccolo rispetto al segnale da misurare, lo strumento va in *overload*, cioè "satura", con il rischio di danneggiarsi (in particolare il tester analogico a lancetta); (ii) se il fondo scala è (troppo) grande, la misura perde di significatività. Nella quasi totalità dei casi, esaminando il circuito sotto analisi e conoscendone a priori alcune caratteristiche si può stabilire il valore massimo che una data grandezza può avere e quindi scegliere di conseguenza la portata. Nel caso in cui questa operazione preliminare sia difficile da eseguire, conviene sicuramente partire da portate alte per poi scendere progressivamente fino ad avere una misura con il massimo della significatività.
3. In linea di principio, oltre alla portata bisogna anche determinare preliminarmente il "segno" della grandezza misurata (cioè, nella pratica, il verso di percorrenza delle correnti all'interno dello strumento o il segno delle differenze di potenziale a esso

applicate). La scelta del segno è critica per lo strumento analogico, dato che la sua lancetta può muoversi solo in un verso (e si può danneggiare se forzata a muoversi in verso opposto). Essa è meno rilevante nel caso del tester digitale, dato che questo strumento ha la possibilità di regolarsi automaticamente sul segno delle grandezze misurate (il segno è indicato sul display).

- Collegare lo strumento al circuito, attraverso i fili di cui è dotato, in modo *corretto*. C'è per esempio una differenza sostanziale tra il collegamento per la misura di tensioni (il tester va collegato "in parallelo" ai punti rispetto ai quali si vuole misurare la d.d.p.) e il collegamento per la misura di correnti (il tester va collegato "in serie" rispetto al ramo di cui si vuole misurare l'intensità di corrente).

Qualche suggerimento generico per la configurazione preliminare degli strumenti:

- state attenti ai colori. Per esempio, nel tester digitale in uso in laboratorio, le scritte in azzurro sul pannello si riferiscono a grandezze in continua (a cui si fa riferimento con DC, ovvero con il simbolo " $=$ "), quelle in bianco a grandezze in alternata (AC, ovvero " \sim "). Nel tester analogico, invece, i colori di DC e AC sono rispettivamente nero e rosso.
- Sempre parlando di colori, ricordate che la convenzione generale per i circuiti elettrici prevede che rosso e nero si riferiscano a punti che si trovano a potenziale rispettivamente maggiore e minore (il rosso è il *positivo*, il nero è il *negativo*). Per esempio, rosse e nere sono le boccole di ingresso del tester digitale.
- State attenti a capire cosa significano le indicazioni accanto alle boccole: molte boccole hanno diverse funzioni (la scelta dipende dalla configurazione di *entrambe* le boccole collegate ai due fili in uso per la misura). In particolare la boccola indicata con COM nel tester digitale e con " $=$ " (in genere con una linea tratteggiata) è a comune per varie funzioni. Essa deve trovarsi a potenziale più basso dell'altra.

Il collegamento con il circuito sotto analisi deve *sempre* avvenire con *due* fili. Il motivo è ovvio: se si vuole misurare una tensione, in quanto *differenza* di potenziale essa può essere determinata solo se il "potenziale" di due punti distinti del circuito viene riportato allo strumento. Analogamente, se si vuole misurare una intensità di corrente è necessario che la corrente *entri nello* (attraverso un filo) ed *esca dallo* (attraverso l'altro filo) strumento. La norma generale è che *tutte* le misure di grandezze, o segnali elettrici, richiedono due fili di collegamento per essere effettuate. Qualche volta, questi fili usati per collegare lo strumento di misura al circuito prendono il nome di *puntali*. I puntali sono propriamente delle punte conduttrici montate su supporti isolanti che spesso si usano

con i tester per collegarsi a punti specifici di un circuito. Nelle vostre esperienze, voi generalmente userete dei fili che terminano con delle banane, o spinotti, standard (4 mm dia). State attenti: per collegarsi alle boccole del tester analogico occorre usare banane specifiche, di diametro minore (nella custodia del tester troverete cavetti con intestati i due tipi di banane, standard e piccolo).

Per quanto riguarda i collegamenti al circuito sotto analisi:

- nella misura di tensione, lo strumento deve "sentire" la d.d.p. tra due punti di un qualche circuito. La misura non richiede di modificare fisicamente il circuito. Se per esempio vi si chiede di misurare la d.d.p. ai capi di un componente (detta anche *caduta di potenziale*, o caduta di tensione), lo strumento di misura, opportunamente configurato come voltmetro, va collegato *in parallelo* al componente stesso.
- Per la misura di intensità di corrente, *tutta* la corrente da misurare deve fluire all'interno dello strumento. Questo richiede di interrompere il circuito nel ramo all'interno del quale volete misurare la corrente e collegare lo strumento, opportunamente configurato come amperometro, *in serie* al ramo stesso.
- Come chiariremo nel seguito, la misura di resistenza, che gli strumenti eseguono sempre e solo in corrente continua, è in qualche modo "derivata" da quelle di tensione o corrente. Per la misura di resistenza, per esempio di un componente (resistivo) o una combinazione di componenti, si deve fare in modo che esso sia collegato *da solo* allo strumento opportunamente configurato come ohmetro. In altre parole, per misurare la resistenza di un resistore dovete *scollegare* il resistore dal circuito in cui si trova eventualmente montato, e collegarlo da solo allo strumento.

III. ANALOGICO VS DIGITALE

Gli aggettivi *analogico* e *digitale*, quando riferiti a strumenti di misura, sottolineano la capacità di fornire letture idealmente continue (analogico) o discrete (digitale). Uno strumento a lancetta, in cui la lancetta può in linea di principio assumere una qualsiasi posizione angolare nell'intervallo consentito, è, ovviamente, uno strumento analogico. Uno strumento dotato di un display numerico è, ovviamente, uno strumento digitale.

I modelli per le due tipologie di strumenti possono differire tra loro, visto che differenti sono i meccanismi di operazione. Qui nel seguito ci occuperemo in particolare del tester analogico, di cui siamo in grado di spiegare quasi tutto.

A. Tester analogico

Il tester analogico che userete (ICE-680R), vanto mondiale di un'industria italiana di qualche decennio fa, si basa su uno strumento che ha una lancetta che si muove sopra una scala graduata (tante scale graduate, con la pratica riuscirete ad associare ogni scala alla misura che state facendo). Per le scale di tensione continua, si hanno in totale 50 tacchette. Il fondo scala della misura è indicato in prossimità di una delle boccole di ingresso. *Per interpretare la lettura, basta ricordare che ogni tacchetta corrisponde a 1/50 del valore di fondo scala.* Sul quadrante è riportato anche uno specchio che serve per minimizzare gli errori di parallasse, e in corrispondenza dell'asse della lancetta c'è una vite, che dovrebbe essere regolata spesso per posizionare sullo zero (entro l'errore di parallasse) la lancetta in condizioni di riposo.

La lancetta è collegata meccanicamente al perno di un telaietto che supporta una bobina di filo conduttore. Il telaietto, cioè la bobina, può ruotare attorno al perno con attrito molto basso; la rotazione è contrastata da un momento di forze elastiche prodotto da una molla che lavora a torsione. Il sistema meccanico ha anche una certa costante tempo di smorzamento, che serve per evitare che la lancetta oscilli indefinitamente e che contribuisce a determinare la prontezza dello strumento. Il telaietto, e quindi la bobina, sono inseriti all'interno di un campo magnetico esterno e costante, generato dalle espansioni di un magnete permanente. Semplificando, facendo scorrere della *corrente* nella bobina si ha che su di essa, e dunque sul telaietto, agisce un momento di forze *proporzionale all'intensità di corrente* e dipendente dal seno dell'angolo compreso tra campo magnetico e direzione ortogonale alla bobina. Questo momento di forze fa ruotare il telaietto, e dunque la bobina, che è detta bobina mobile, fino a una posizione determinata dall'equilibrio con il momento della forza elastica della molla di torsione. La deflessione della lancetta è quindi proporzionale all'*intensità di corrente* che fluisce nella bobina. Lo strumento a lancetta è allora, fondamentalmente, un misuratore di corrente, detto *galvanometro* o (micro)amperometro. La Fig. 1 ne riporta uno schema semplificato, con viste di profilo e dall'alto.

Due osservazioni molto rilevanti:

1. per come è costruito (la bobina è un avvolgimento di filo elettrico), lo strumento ha inherentemente una *resistenza interna* r non nulla, data almeno dalla resistenza del filo che costituisce l'avvolgimento. Un modello dello strumento, in questa configurazione di base che stiamo considerando, può quindi essere quello rappresentato in Fig. 2(a), dove un amperometro "ideale" (privo di resistenza, in questa accezione) si trova in serie alla resistenza r .
2. Lo strumento è in realtà un trasduttore, dato che converte la misura della grandezza elettrica (corrente) nello spostamento della lancetta. La traduzione di una grandezza in un'altra richiede di

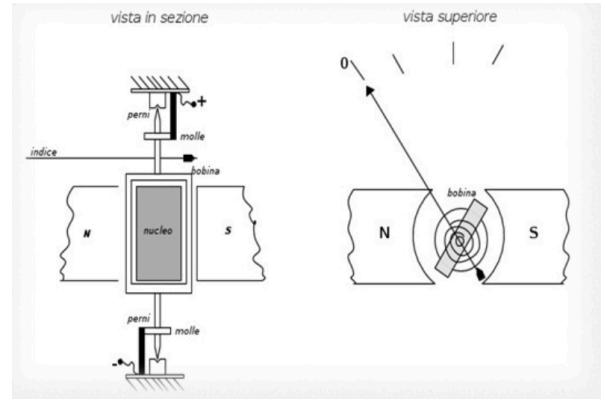


Figura 1. Rappresentazione schematica di un galvanometro o (micro)amperometro, con viste di profilo ("in sezione") e dall'alto ("superiore").

determinare un coefficiente di ragguaglio, o di *calibrazione*. Esso viene normalmente individuato dal costruttore, che si fa anche carico di stabilire l'accuratezza della calibrazione stessa.

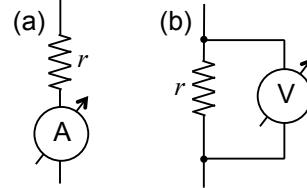


Figura 2. Rappresentazione schematica di un amperometro (a) e di un voltmetro (b) reali; gli strumenti ideali sono rappresentati dal circoletto con freccina, e si intendono qui dotati di resistenza nulla (amperometro) o infinita (voltmetro); r rappresenta la resistenza interna dello strumento reale.

B. Tester digitale

In linea di massima, e senza esaminare casi particolari, le grandezze elettriche che si misurano in un laboratorio sono "continue", nel senso qui di "non discrete". Infatti nelle comuni misure elettriche, in cui sono coinvolte grandi quantità di carica, è estremamente difficile rendersi conto della natura discreta della carica elettrica (la corrente di $1 \mu\text{A}$, che è piccola per i nostri standard, corrisponde al flusso di $10^{12} - 10^{13}$ elettroni al secondo). In uno strumento digitale viene eseguita una conversione da analogico e digitale, cioè la grandezza da misurare viene *digitalizzata* e convertita in un numero a cui poi viene attribuita una unità di misura a seconda della configurazione dello strumento. Tratteremo i meccanismi di base della digitalizzazione, che al giorno d'oggi è una tecnica comunissima in tanti ambiti, in altra sede.

Qui ci limitiamo a sottolineare che un tester digitale può essere considerato, in termini fondamentali e in

prima battuta, come un *misuratore di d.d.p.*. Poiché lo strumento deve poter eseguire tante diverse funzioni e operare su tante diverse scale, è evidente che oltre al digitalizzatore vero e proprio esso sarà costituito da molta circuiteria elettronica, finalizzata per esempio ad amplificare o attenuare i segnali in modo da permettere di cambiare scala o tipologia di misura. Tutto questo fa sì che:

1. per come è costruito (non sappiamo bene come), lo strumento ha inherentemente una *resistenza interna* r non infinita, che può essere considerata come montata in parallelo a un voltmetro “ideale” (in questa accezione, dotato di resistenza infinita), come rappresentato in Fig. 2(b);
2. naturalmente anche in questo caso lo strumento è in realtà un trasduttore, e quindi la misura si basa sulla calibrazione, che deve essere fornita dal costruttore assieme alla sua incertezza.

C. Misure su diverse portate e di diverse grandezze

La multifunzionalità dei tester, così come la possibilità di selezionare diverse portate, è in gran parte conseguenza diretta dell'applicazione della cosiddetta legge di Ohm. Poiché non siamo in grado di descrivere nei dettagli il funzionamento di uno strumento digitale, per capire questa affermazione facciamo riferimento alla costruzione del solo tester analogico. La Fig. 3 mostra lo schema circuitale interno, debitamente semplificato, dello strumento in uso in laboratorio. Se facciamo riferimento allo schema (a), possiamo notare come un'opportuna configurazione di resistenze, montate internamente allo strumento e connesse a diverse boccole di entrata, renda possibile usare il galvanometro, che di per sé è un misuratore di corrente, per la misura di differenze di potenziale.

Infatti supponiamo di applicare una differenza di potenziale ΔV tra la boccola marcata con V in figura (sarebbe la boccola “comune”, quella indicata con “=” sul pannello dello strumento) e la boccola marcata con 2V (che indica un fondo scala di 2 V). Questa d.d.p. provoca una corrente che passa nel circuito costituito dalla serie della resistenza indicata come R_1 con il parallelo dei due rami costituiti dalla resistenza R_2 e dalla serie della resistenza R_3 con il galvanometro (indicato con A e la freccina).

Scriviamo per esteso il sistema di equazioni che permettono di determinare il valore della corrente I_B che passa per il galvanometro. Per semplicità, immaginiamo di trascurare la resistenza interna del galvanometro stesso (essa è piccola, anche se non nulla). Si ha:

$$R_{tot} = R_1 + R_{2//}R_3 = R_1 + \frac{R_2R_3}{R_2 + R_3} \quad (1)$$

$$\Delta V = R_{tot}I_{tot} = R_{tot}(I_A + I_B) \quad (2)$$

$$\frac{I_A}{I_B} = \frac{R_3}{R_2}, \quad (3)$$

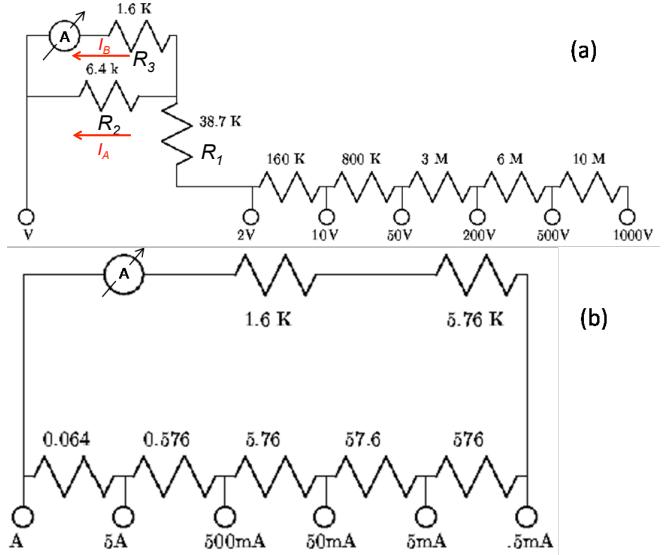


Figura 3. Schema semplificato del tester analogico ICE-680R in configurazione per la misura di d.d.p. (a) e di intensità di corrente (b).

dove i vari simboli hanno un significato ovvio. Risolvendo si ha

$$I_B = \frac{R_2}{R_2 + R_3} \frac{\Delta V}{R_{tot}} = \frac{R_2}{R_1(R_2 + R_3) + R_2R_3} \Delta V. \quad (4)$$

Questa equazione dimostra che c'è proporzionalità diretta tra il valore della tensione ΔV e la corrente I_B che, circolando nel galvanometro, determina la deflessione della lancetta. Quindi da un amperometro abbiamo costruito un voltmetro, e il tutto è stato dimostrato usando la legge di Ohm. Un'ispezione quantitativa dei valori delle varie resistenze (indicati in modo barbaro in figura) mostra che per $\Delta V = 2$ V si ha $I_B = 40 \mu\text{A}$, valore che corrisponde alla massima deflessione della lancetta, e quindi al fondo scala.

Se si vuole cambiare fondo scala, cioè la portata dello strumento, è sufficiente aggiungere delle resistenze in serie. Per esempio, se invece della boccola marcata con 2V si usa quella marcata con 10V, si ha un'ulteriore resistenza in serie a R_1 . Usando i valori riportati in figura si vede facilmente che in questo caso il fondo scala del galvanometro (cioè la condizione $I_B = 40 \mu\text{A}$) corrisponde a una d.d.p. applicata tra le boccole considerate pari a $\Delta V = 10$ V (e infatti il fondo scala prescelto per la misura di tensione con quella scelta delle boccole è proprio questo). Quindi, usando opportuni elementi puramente resistivi posti all'interno dello strumento, è possibile modificare a volontà la sua portata.

La modifica della portata può essere facilmente ottenuta anche nel caso di misure di corrente. Facciamo riferimento alla Fig. 3(b) per capire come questo sia realizzato nel tester analogico in uso in laboratorio. Supponiamo allora di inviare una corrente di intensità I facendola entrare per la boccola marcata .5mA e facendola uscire da

quella marcata con A. Questa corrente viene ripartita tra due rami: quello in alto è costituito dalla serie di due resistenze con il galvanometro, quello in basso è costituito dalla serie di cinque resistenze. Se si fanno i conti, si vede facilmente che per questa scelta di boccole la corrente $I = 0.5 \text{ mA}$ corrisponde al passaggio di $40 \mu\text{A}$ attraverso il galvanometro. Dunque il fondo scala per questa scelta di boccole è 0.5 mA . Se invece della boccola 5mA si invia la corrente attraverso la boccola 5mA si modifica la configurazione del circuito (guardate lo schema). Se si fanno i conti, si vede facilmente come in questa nuova configurazione la deflessione massima del galvanometro corrisponda a una corrente di 5 mA . Dunque la portata è stata modificata di un fattore noto usando gli elementi resistivi interni allo strumento.

Del tester digitale non siamo in grado di esaminare lo schema circuitale, dato che esso comprende parecchia circuiteria elettronica che non conosciamo (almeno per ora). Tuttavia abbiamo già stabilito che esso si comporta, di base, come un voltmetro. Dunque usando schemi non troppo diversi, almeno concettualmente, rispetto a quelli di Fig. 3 possiamo immaginare di poterne convertire il funzionamento in amperometro e anche di poter cambiare a volontà la portata.

Infine spendiamo due parole sulla misura di resistenza. La legge di Ohm definisce la resistenza come il rapporto tra la caduta di potenziale ai capi del componente che si sta misurando e la intensità di corrente che lo attraversa. Dunque la resistenza può essere determinata avendo o un misuratore di corrente e una sorgente di d.d.p. nota, applicata ai capi del componente, oppure un misuratore di d.d.p. e una sorgente di corrente nota, da far scorrere attraverso il componente. Queste due possibilità sono quelle utilizzate rispettivamente nel tester analogico e in quello digitale. Nel caso del tester analogico, la misura della resistenza viene effettuata utilizzando una batteria interna (normalmente scarica, e quindi inutilizzabile!) che viene collegata ai capi del componente di resistenza incognita. Nel collegamento, si fa in modo che il galvanometro e le resistenze necessarie per ottenere la corretta portata nella misura di corrente si vengano a trovare in serie con il componente sotto misura. La resistenza dipende in modo inversamente proporzionale con la corrente che fluisce nel componente e che dunque è letta dallo strumento. Infatti la scala per la lettura delle resistenze ha il suo zero al fondo scala e presenta un andamento visibilmente nonlineare.

A causa della complessità interna dei circuiti coinvolti, un'analisi di questo tipo non può essere realizzata per il tester digitale. Tuttavia possiamo ipotizzare un modello in cui il voltmetro che costituisce la base dello strumento viene utilizzato per misurare la d.d.p. ai capi del componente sotto misura, in cui viene fatta passare una corrente di intensità nota prodotta da un opportuno circuito elettronico. In una esercitazione pratica futura cercheremo di verificare la validità di questa ipotesi.

IV. RESISTENZE INTERNE DEGLI STRUMENTI

In questa sezione trattiamo un argomento della massima importanza, da tenere in debito conto *ogni volta* che si eseguono misure di segnali elettrici, a prescindere dallo strumento.

Supponiamo di voler misurare una corrente con uno strumento reale, modellato secondo la Fig. 2(a). Tale strumento ha una resistenza interna non nulla che, nella misura di corrente, viene a trovarsi in serie al circuito sotto analisi. Questa resistenza interna “perturba” (cioè influenza) il funzionamento del circuito esterno, ed è evidente che questa “perturbazione” è tanto maggiore quanto più grande è il valore della resistenza interna. Infatti la resistenza interna dello strumento viene a trovarsi in serie nel ramo di circuito considerato per la misura e quindi modifica l'intensità di corrente che lo attraversa rispetto al caso imperturbato. Dunque *un amperometro reale approssima il comportamento ideale quanto più piccola è la sua resistenza interna*.

Supponiamo ora di voler misurare una tensione con uno strumento reale, modellato secondo la Fig. 2(b). Tale strumento ha una resistenza interna non infinita che, nella misura di tensione, viene a trovarsi in parallelo al componente, o alla composizione di componenti, ovvero al ramo del circuito, ai capi del quale vogliamo misurare la d.d.p.. Questa resistenza interna “perturba” il funzionamento del circuito esterno, ed è evidente che questa “perturbazione” è tanto maggiore quanto più piccolo è il valore della resistenza interna. Infatti c'è una frazione non nulla di corrente che, invece di attraversare il ramo sotto analisi come nel caso imperturbato, fluisce nello strumento. Dunque *un voltmetro reale approssima il comportamento ideale quanto più grande è la sua resistenza interna*.

I dati relativi alle resistenze interne, o informazioni che permettono di determinarle, sono sempre specificati nei manuali degli strumenti, a cui si rimanda per ulteriori dettagli. Diamo qui un cenno dei valori tipici cominciando dall'impiego degli strumenti analogico e digitale come voltmetri, che è l'applicazione più frequente e significativa.

- Per il tester analogico, lo schema di Fig. 3(a) mostra chiaramente che la resistenza interna (che abbiamo prima chiamato R_{tot}) dipende dalla portata selezionata. Essa aumenta all'aumentare del fondo scala. Facendo due conti, si vede che $r = 20 \text{ kohm}/V_{fs}$, dove V_{fs} è il *fondo scala* (in V) della portata selezionata. Come rule of thumb, si può affermare che gli effetti della resistenza interna, ovvero della perturbazione, sono trascurabili se $r >> R$, con R resistenza ai capi della quale si sta misurando la tensione, ovvero resistenza del ramo considerato per la misura. Le tensioni tipiche misurate in laboratorio sono di qualche V, per cui i fondo scala impiegati sono tali che la resistenza interna è di

alcune decine di kohm, un valore non sempre trascurabile rispetto a quello delle resistenze di uso comune nelle esperienze pratiche. Dunque *il tester analogico non è quasi mai la scelta ottimale per la misura di tensioni.*

- Nel tester digitale la resistenza interna è dominata da quella del circuito di ingresso che è in genere realizzato con transistor a effetto di campo. In questi circuiti la resistenza di ingresso è tipicamente molto alta. Nel tester digitale in uso in laboratorio si ha $r \simeq 10$ Mohm, indipendentemente dalla portata selezionata. Dunque *il tester digitale produce quasi sempre perturbazioni trascurabili nella misura di tensioni* ed è quindi preferibile a quello analogico.
- Per la misura di correnti con il tester analogico facciamo riferimento alla Fig. 3(b). Si vede chiaramente che anche in questo caso la resistenza interna dipende dalla scala (essa può essere determinata svolgendo i calcoli). Tradizionalmente, però, invece che attraverso la resistenza interna l'effetto di perturbazione provocato dall'inserzione dello strumento in serie al circuito sotto analisi viene caratterizzato attraverso la *caduta di potenziale per inserzione* per lettura a fondo scala, $\Delta V_{\text{ins},fs}$. Questa caduta di potenziale è (generalmente) *costante*, cioè indipendente dalla portata selezionata, come si può facilmente verificare usando la legge di Ohm. Tuttavia essa dipende linearmente dalla corrente effettivamente misurata; i manuali normalmente forniscono la caduta di potenziale corrispondente a una lettura a fondo scala. Per il tester analogico, si ha $\Delta V_{\text{ins},fs} \simeq 300$ mV (tranne che per la portata con fondo scala $50 \mu\text{A}$, per la quale $\Delta V_{\text{ins},fs} \simeq 100$ mV), per il digitale $\Delta V_{\text{ins},fs} \simeq 200$ mV. Non è sempre immediato stabilire a priori se queste cadute di potenziale corrispondono a perturbazioni trascurabili, o meno, dei circuiti sotto analisi, ma *entrambi i tester (analogico e digitale) perturbano in genere allo stesso modo* i circuiti quando usati come amperometri. Dunque, non c'è sostanziale preferenza per l'uno o l'altro se si devono eseguire misure di intensità di corrente.

Un'ultima considerazione praticamente rilevante: se si usa *erroneamente* un multimetro configurato come amperometro per la misura di tensioni, è evidente che si possono provocare *seri danni* sia allo strumento che al circuito sotto analisi. Infatti la bassa resistenza interna corrisponde in pratica, nella maggior parte dei casi, a provocare un corto circuito. Anche la situazione inversa (uso di un multimetro configurato come voltmetro per misurare correnti) va attentamente evitata.

A. Esempi

Per chiarire quali siano gli effetti della perturbazione creata dall'uso degli strumenti di misura, facciamo riferimento a due casi specifici di circuiti comprendenti delle resistenze e un generatore di d.d.p.. Per semplicità, in questi esempi supponiamo di disporre di un generatore di d.d.p. *ideale*, cioè in grado di fornire la d.d.p. assegnata (V_0 , in questi esempi) a prescindere dal circuito che vi è collegato.

Il primo esempio è un partitore di tensione costituito dalle resistenze R_1 e R_2 , in cui immaginiamo di voler misurare la tensione V_1 ai capi di R_1 [Fig. 4(a)]. In assenza dello strumento ci si attenderebbe una tensione

$$V_{1,\text{att}} = R_1 I = \frac{R_1}{R_1 + R_2} V_0 , \quad (5)$$

con ovvio significato dei simboli. La presenza del voltmetro reale modifica il circuito, dato che la sua resistenza interna r viene a trovarsi *in parallelo* a R_1 . Dunque il partitore di tensione è costituito dalla serie di R_2 con il parallelo $R_{1//r} = R_1 r / (R_1 + r)$. Questo modifica l'intensità di corrente I' che circola nel circuito (ovvero nella maglia). Di conseguenza la tensione misurata è

$$V_{1,\text{mis}} = R_{1//r} I' = \frac{r R_1}{r R_1 + r R_2 + R_1 R_2} V_0 \quad (6)$$

$$= \frac{R_1}{R_1 + R_2 + R_2/\xi} , \quad (7)$$

con $\xi = r/R_1$. Evidentemente $V_{1,\text{mis}} \rightarrow V_{1,\text{att}}$ per $r \gg R_1$, soddisfatta di certo per r tendente a infinito.

Facciamo un esempio numerico, in cui poniamo $V_0 = 4.5$ V, $R_2 = 20$ kohm, $R_1 = 10$ kohm. Il partitore di tensione così creato è atteso fornire $V_{1,\text{att}} = 1.5$ V. Nel caso di uso del tester analogico, immaginando di usare la portata con fondo scala 2V a cui corrisponde $r = 40$ kohm, ovvero $\xi = 4.0$, si ha $V_{1,\text{mis}} = 1.3$ V (differente dal valore atteso per oltre il 10%). Nel caso invece di tester digitale, che ha $r \simeq 10$ Mohm, si ottiene facilmente $V_{1,\text{mis}} = V_{1,\text{att}}$.

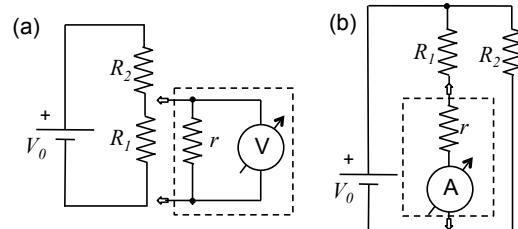


Figura 4. Partitore di tensione (a) e di corrente (b) considerati nel testo. I box tratteggiati indicano i modelli di strumenti di misura reali.

Il secondo esempio è un partitore di corrente costituito dalle resistenze R_1 e R_2 in cui immaginiamo di voler misurare l'intensità di corrente I_A che scorre per il ramo in cui si trova R_1 [Fig. 4(b)]. In assenza dello strumento e

considerando ideale il generatore, ci si attenderebbe una corrente

$$I_{A,att} = \frac{V_0}{R_1} = \frac{R_2}{R_1} I_B , \quad (8)$$

dove I_B è la corrente che passa per il ramo costituito da R_2 . Inserendo lo strumento, che va ovviamente collegato in serie con R_1 , si misura una corrente

$$I_{A,mis} = \frac{V_0}{r + R_1} = \frac{R_2}{r + R_1} I_B . \quad (9)$$

Si vede subito che $I_{A,mis} \rightarrow I_{A,att}$ per $r \rightarrow 0$ (in queste condizioni è anche soddisfatta l'equazione del partitore di corrente espressa dall'ultimo membro di Eq. 8). Introducendo la caduta di potenziale per inserzione $\Delta V_{ins} = r I_{A,mis}$, si ha anche

$$I_{A,mis} = \frac{V_0 - \Delta V_{ins}}{R_1} , \quad (10)$$

che evidentemente si riconduce a Eq. 8 se $\Delta V_{ins} \rightarrow 0$.

Anche qui facciamo un esempio numerico usando gli stessi valori delle varie grandezze menzionate prima. Immaginiamo di usare il tester digitale configurato come amperometro con fondo scala 2mA, in modo da avere una lettura significativa. Si ottiene $I_{A,att} = 0.45$ mA, che corrisponde all'incirca a un quarto del fondo scala. Di conseguenza la caduta di potenziale per inserzione vale $\Delta V_{ins} \approx \Delta V_{ins,fs}/4 \approx 68$ mV. Questo porta a $I_{A,mis} \approx 0.44$ mA, diversa dal valore atteso per circa il 2% (potete provare a determinare il risultato nel caso di uso del tester analogico).

V. LETTURA DELLE SCALE

Questa sezione si occupa di un aspetto assolutamente banale, ma tuttavia molto rilevante dal punto di vista pratico, cioè la lettura della scala, sia per il tester analogico che per il digitale.

La lancetta del tester analogico si muove sopra un gran numero di scale graduate, che si riferiscono alle varie grandezze misurabili e alle varie portate che possono essere selezionate. Dato che in questa nota ci occupiamo in particolare dell'uso del tester per la misura di tensioni e intensità di corrente continue (la misura di resistenza non verrà mai eseguita con il tester analogico), possiamo restringerci a considerare le quattro scale in nero accanto a cui sono posti i simboli V - mA =. Queste scale si riferiscono allo stesso sistema di 50 tacchette equispaziate (le scale sono lineari rispetto alla deflessione della lancetta). Come già affermato, il modo più semplice, ma non immediato, per leggere la scala consiste nel considerare il fondo scala prescelto (indicato accanto a una delle boccole che vengono usate) e dividerlo per 50: ogni tacchetta corrisponde infatti a 1/50 del fondo scala, e contando le tacchette a partire dallo zero (deflessione nulla) si ottiene la lettura. Un modo più immediato consiste nell'usare

i numeri riportati accanto alle tacchette maggiori, che si riferiscono a fondi scala pari a 250, 200, 50, 10. Per esempio, se stiamo facendo una misura di tensione e il fondo scala prescelto è $V_{fs} = 2$ V, allora potremo considerare la serie di numeri che termina con 200 e dividerli mentalmente per 100, oppure la serie che termina con 10 e moltiplicarla per 0.2, allo scopo di ottenere una scala che va da 0 a 2 V.

Queste difficoltà di lettura non esistono nel caso del tester digitale, dove il display numerico (detto, per il modello in uso in laboratorio, DM-3900, a 3 cifre e "mezza", la "mezza" cifra è quella iniziale e può essere assente oppure valere 1) permette di conoscere immediatamente la lettura. Tuttavia con questo modello specifico di tester bisogna ancora porre delle attenzioni per indovinare la scala, cioè l'unità di misura della lettura. Come si vede dalle indicazioni riportate accanto alla manopola del commutatore rotante, i fondi scala impostabili corrispondono a grandezze tipo 200 mV, 2 V, 20 V, 200 V, etc., oppure 20 μ A, 200 μ A, 2 mA, etc. (cominciano tutte per 2, salvo qualche eccezione). Infatti il display impiegato può al massimo fornire l'indicazione numerica 1999. Usando il tester si nota anche che sul display compare un punto decimale la cui posizione dipende dalla portata selezionata. Supponiamo di fare una misura di corrente continua avendo selezionato la portata con fondo scala 200 μ A: il punto decimale si troverà tra la penultima e l'ultima cifra sulla destra, per cui la lettura potrà andare da 0 a 199.9 μ A. Supponiamo ora di fare una misura di tensione avendo selezionato la portata con fondo scala 2 V: il punto decimale si troverà tra la prima ("mezza") cifra e la seconda, e la lettura potrà andare da 0 a 1.999 V.

Altra piccola osservazione: se il display indica la sola "mezza" cifra iniziale, cioè un 1. non seguito da nulla, allora vuol dire che siete in condizioni di overload, cioè che la grandezza che volete misurare ha un valore che eccede il fondo scala della portata prescelta.

VI. INCERTEZZA

Determinare l'incertezza di una misura è sempre compito di chi esegue la misura. Ovviamente la determinazione deve essere basata su considerazioni valide dal punto di vista fisico. Sempre in termini molto generali, in ogni strumento di misura possono essere spesso identificate (almeno) due sorgenti di "incertezza", o di errore. Infatti, in maniera manichea, si può spesso distinguere fra *incertezza di lettura*, generalmente dominata da effetti stocastici, e *incertezza di calibrazione*, generalmente associata soprattutto a effetti non stocastici, detti anche (impropriamente) sistematici.

In modo sintetico, convenzionale e grossolanamente possiamo affermare che entrambi gli strumenti (analogico e digitale) hanno un'incertezza di calibrazione, a cui, altrettanto convenzionalmente e grossolanamente, attribuiamo carattere prevalente *non stocastico*. Essa è dovuta ai tanti

motivi che concorrono a determinare il fattore di calibrazione del processo di trasduzione coinvolto. Per esempio, facendo riferimento al tester analogico, la calibrazione è necessaria nello stabilire la relazione tra corrente che passa nella bobina mobile e deflessione della lancetta (fisicamente essa dipende da rigidità delle molle, intensità del campo magnetico statico applicato, geometria effettiva della bobina, etc.). Inoltre, quando si cambia portata o tipo di grandezza misurata usando sistemi di resistenze, come visto sopra, la calibrazione è necessaria per stabilire il comportamento di questi sistemi. In linea di principio, la calibrazione, che in genere si effettua tramite confronto con uno *standard*, potrebbe essere definita con un'accuratezza paragonabile a quella con cui è definito lo standard stesso. Tuttavia questa operazione, che va fatta individualmente strumento per strumento, non è stata eseguita per gli apparecchi in uso in laboratorio (non viene mai eseguita per tester di quel tipo), per cui il costruttore esprime una sorta di tolleranza sulla calibrazione che si traduce, di fatto, in un'incertezza, detta appunto di calibrazione. Una situazione analoga, anzi, più complicata, si verifica con il tester digitale, dove il digitalizzatore e tutta l'elettronica di contorno devono essere opportunamente calibrati.

Osservate che il carattere manicheo dell'attribuzione non stocastica a questa fonta di incertezza cozza con la circostanza che alcuni dei fattori che la determinano sono a loro volta affetti da variazioni stocastiche legate, ad esempio, alle condizioni di operazione (temperatura, pressione, umidità, etc.). Tuttavia in genere queste variazioni producono effetti piccoli e il carattere dominante resta non stocastico.

Accanto all'incertezza di calibrazione per entrambi gli strumenti possiamo individuare un'incertezza di lettura, da intendere in senso lato, che invece ha carattere prevalentemente stocastico. Nel tester analogico tale incertezza dipende effettivamente dalla lettura della deflessione della lancetta e può corrispondere alla minima variazione della sua posizione resa possibile dalla misura a occhio, cioè, per esempio, \pm "mezza tacchetta". Il per esempio sta ad indicare che la stima dell'incertezza di lettura può dipendere da fattori quali l'abilità del lettore, il fatto che la lancetta non sia piegata o abbia un certo spessore, la circostanza che essa non vibri per qualche motivo, etc.. Nel tester digitale, invece, all'incertezza di lettura intrinseca del dato, che è facilmente individuabile in \pm "mezza cifra meno significativa", o, con una terminologia che useremo spesso, ± 1 *digit*, occorre normalmente aggiungere un contributo che dipende dai complessi meccanismi coinvolti nella conversione analogico/digitale.

Infine si rimanda all'Appendice per un metodo, a voi già perfettamente noto, che permette di convertire le incertezze di lettura a deviazioni standard, come può essere utile in qualche circostanza.

A. Calibrazione vs lettura

L'incertezza di calibrazione, che in genere dipende da portata e grandezza misurata, è sempre specificata nel manuale dello strumento; nel caso del tester digitale, il manuale riporta anche, portata per portata, l'incertezza di lettura, tipicamente espressa in digit. Normalmente, invece, l'incertezza di calibrazione è espressa come un *errore relativo* (percentuale). Le due sorgenti di incertezza (calibrazione e lettura) possono essere ritenute convenzionalmente indipendenti. Di conseguenza l'errore complessivo sulla misura si ottiene convenzionalmente *sommando in quadratura* i due termini di incertezza [1]. Per strumenti della categoria di quelli in uso in laboratorio l'effetto complessivo è tipicamente "notevole": molto spesso le misure di grandezze elettriche sono caratterizzate da errori che sovraстimano ampiamente la deviazione standard (stocastica) della grandezza misurata o la sensibilità dello strumento. Ve ne potete facilmente rendere conto provando a stabilire la deviazione standard sperimentale di una certa misura, per esempio della d.d.p. prodotta da un generatore: spesso vedrete che le fluttuazioni della lettura (eventualmente visibili a occhio nelle cifre ballerine) sono molto minori della barra di errore complessiva attribuita alla misura stessa, in una misura dove la deviazione standard potrebbe essere dovuta anche a fluttuazioni della d.d.p. erogata dal generatore.

Questa circostanza può essere fastidiosa in alcuni casi, per esempio perché impedisce di eseguire test di significatività del best-fit. Quindi potrebbe essere utile, in alcuni casi *opportunamente selezionati e discussi*, trascurare l'incertezza di calibrazione, che in genere è il fattore dominante nella barra di errore. Vedremo caso per caso se e quando questa operazione può essere eseguita. Per il momento, limitiamoci a dichiarare alcune situazioni in cui l'incertezza di calibrazione *non* può essere trascurata:

1. quando la lettura, anche se fatta con lo stesso strumento e nelle stesse condizioni operative, richiede di cambiare portata. Infatti il fattore di calibrazione è specifico per ogni portata e dunque trascurare la relativa incertezza può condurre a sottostimare l'errore.
2. Quando si devono combinare misure di grandezze diversamente dimensionate in una misura "indiretta". Per esempio, questo è il caso di un best-fit dei dati di resistenza e tensione (ai capi della resistenza) usato per determinare la validità di un modello basato sulla legge di Ohm. Poiché resistenza e tensione, magari misurati con lo stesso strumento (lo stesso tester digitale), hanno dimensioni diverse, e dunque sfruttano diversi fattori di calibrazione, trascurare la relativa incertezza può condurre anche in questo caso a sottostimare l'errore.

B. Esempi

Per familiarizzare con la determinazione delle incertezze facciamo riferimento a qualche caso pratico. Allo scopo di contestualizzare la rilevanza dell'argomento, ricordate sempre che l'errore è una *stima*: di conseguenza, molto spesso può essere ragionevole valutarlo in modo “rapido”, senza sforzarsi di fornire un valore con tante cifre significative che sono spesso del tutto inutili. In linea di massima è fortemente sconsigliato, e forse anche vietato, servirsi di eccessivi ausili per la determinazione delle incertezze, per esempio programmini di computer che svolgono il compito in maniera egregia, ma inutile: meglio usare la testa (in genere più che sufficiente), la matita (quasi sempre inutile) o, al limite, la calcolatrice (generalmente inutile).

Immaginiamo di leggere la d.d.p. V_0 prodotta da un generatore usando il tester digitale. Supponiamo che il generatore sia uno di quelli in uso in laboratorio, in cui $V_0 \sim 5$ V. Predisponiamo il tester sulla portata con fondo scala 20V (il fondo scala inferiore sarebbe al di sotto del valore che ci aspettiamo di leggere, quello superiore peggiorerebbe la significatività della misura) e assumiamo di vedere sul display la lettura $V_0 = 4.97$ V (il display ha tre cifre e “mezza”, dunque il massimo numero di cifre significative ottenibili in questo caso è tre). Il manuale dello strumento, disponibile in rete e in laboratorio, recita che per questa portata e questa grandezza la “precisione” è $\pm 0.5\% \pm 1$ digit. Questi due termini corrispondono proprio all’incertezza di calibrazione e di lettura. L’incertezza di calibrazione produce una (semi)barra di errore $\Delta V_{0,cal} = 0.025$ V (usando una cifra significativa di ridondanza). L’incertezza di lettura produce una (semi)barra di errore $\Delta V_{0,lett} = 0.01$ V (minore di quella di calibrazione). La (semi)barra di errore complessivo vale allora $\Delta V_0 = \sqrt{0.025^2 + 0.01^2}$ V = 0.027 V. Sopprimendo la cifra significativa di ridondanza, si potrà esprimere la misura come $V_0 = (4.97 \pm 0.03)$ V. Notate che, se avessimo scelto (erroneamente) la portata successiva, con fondo scala 200V, la lettura sarebbe stata probabilmente $V_0 = 5.0$ V, con sole due cifre significative. Tenendo conto che questa scala ha la stessa “precisione” di quella usata in precedenza, avremmo ottenuto sempre $\Delta V_{0,cal} = 0.025$ V, ma $\Delta V_{0,lett} = 0.1$ V (superiore in questo caso all’incertezza di calibrazione). Di conseguenza, tenendo conto delle cifre significative, avremmo ottenuto $V_0 = (5.0 \pm 0.1)$ V (la misura è “meno significativa” di quella fatta in precedenza).

Facciamo un altro esempio con la misura di un resistore con quattro anelli colorati, i cui primi tre sono arancione-arancione-rosso. Usiamo il tester digitale configurato come ohmetro con fondo scala 20k e supponiamo che la lettura sia 3.78 (di nuovo con tre cifre significative), che significa, tenendo conto del fondo scala, $R = 3.78$ kohm. Il manuale recita che la “precisione” è, in questo caso, $\pm 0.8\% \pm 1$ digit. Si ottiene allora $\Delta R_{cal} = 0.030$ kohm e $\Delta R_{lett} = 0.01$ kohm, da cui $R = (3.78 \pm 0.03)$ kohm, con errore dominato anche qui dalla calibrazione.

Ancora un altro esempio. Stavolta abbiamo una corrente di intensità $I \sim 15$ mA misurata dal tester analogico. Scegliamo la portata con fondo scala 50mA e supponiamo di osservare che la lancetta si ferma attorno alla diciottesima tacchetta. Questo vuol dire che la lettura è $I = 18.0$ mA. Per le condizioni nelle quali operiamo, stabiliamo che $\Delta I_{lett} = 0.5$ mA, mentre il manuale recita che la “precisione” è $\pm 1\%$ del fondo scala, che individuiamo come incertezza di calibrazione. Di conseguenza $\Delta I_{cal} = 0.5$ mA (pari a quella di lettura) e, in definitiva, $I = (18.0 \pm 0.7)$ mA.

Come rule of thumb, almeno per il tester digitale, l’incertezza di calibrazione prevale quando la misura occupa “gran parte” della scala prevista da una certa portata, condizione in cui, ricordando che la determinazione dell’incertezza è *sempre* una *stima*, si può talvolta trascurare l’incertezza di lettura. D’altra parte è ovvio che questo *non* può essere fatto quando la lettura si avvicina allo zero, dato che un’incertezza nulla *non* è fisicamente accettabile in alcuna circostanza.

APPENDICE: INCERTEZZA E DEVIAZIONE STANDARD

Immaginiamo di eseguire una lettura con lo strumento analogico a lancetta, cioè in condizioni nelle quali sappiamo ben identificare l’incertezza di lettura, e supponiamo che questa valga \pm “mezza tacchetta”. Supponiamo poi di poter trascurare, per un caso miracoloso, l’incertezza di calibrazione, per esempio perché siamo interessati a ricostruire un trend, per cui non ci interessa la conversione della lettura (in “tacchette”) in unità fisiche, oppure perché abbiamo appena eseguito una calibrazione ad hoc con uno standard di elevata accuratezza.

Per come abbiamo costruito l’incertezza, possiamo affermare che *sicuramente* una certa lettura, per esempio 32 “tacchette”, corrisponde a una misura compresa tra 31.5 e 32.5 tacchette, cioè compresa entro una tacchetta attorno al suo valore medio. Secondo quanto stabilito, pur se in maniera rozza e convenzionale, possiamo attribuire a questa incertezza una natura prevalentemente stocastica, per cui possiamo chiederci quanto vale la *deviazione standard* σ corrispondente.

Una possibilità consiste nel supporre che la misura sia rappresentativa di una distribuzione normale (Gaussiana) di valori, per la quale sappiamo che nell’intervallo contenuto in due deviazioni standard (“una sigma”, in linguaggio un po’ gergale) si trova circa il 68% del campione di misure, mentre oltre il 99% sta in un intervallo tre volte più ampio (“tre sigma”). La corrispondenza tra incertezza stabilita e deviazione standard dipende a questo punto dal significato che attribuiamo all’avverbio “sicuramente” usato qui sopra. Possiamo per esempio immaginare che il sicuramente corrisponda al 99% della probabilità, cioè che “pressoché tutte” le misure siano comprese nell’intervallo di una tacchetta, che quindi equivale di fatto a 6σ . Con queste premesse possiamo concludere che la devia-

zione standard è stimabile come un sesto di tacchetta. È evidente che questa affermazione corrisponde in genere a una sottostima della deviazione standard, ed è anche evidente che l'avverbio potrebbe avere altri significati, per esempio perché lo spessore finito della lancetta e delle tacchette rende poco plausibile che il 99% delle misure cada nell'intervallo considerato. Potrebbe quindi essere più sensato accontentarsi di poco più del 95%, che per una distribuzione normale corrisponde a un intervallo di 4σ , per cui la deviazione standard sarebbe stimabile in un quarto di tacchetta.

Un'ulteriore possibilità, che di fatto si rivela particolarmente appropriata nel caso di strumenti digitali, consiste nel supporre che la misura sia rappresentativa di una distribuzione uniforme, ovvero rettangolare con lato di

base pari all'incertezza. Come sapete, la deviazione standard di una distribuzione uniforme è pari a $1/\sqrt{12}$ volte la sua larghezza. Pertanto, immaginando con un po' di difficoltà che anche nell'esempio precedente sia possibile ipotizzare una distribuzione uniforme, la deviazione standard risulterebbe $(1/\sqrt{12}) \simeq (1/3.5)$ tacchette, un valore paragonabile a quanto ottenuto nel caso "due sigma".

Nelle tipiche condizioni delle nostre osservazioni sperimentali, dove l'incertezza di calibrazione gioca un ruolo dominante, la rilevanza del ragionamento di cui sopra è tuttavia modesta, dato che già l'incertezza di lettura è spesso trascurabile rispetto a quella di calibrazione, e quindi considerare la deviazione standard invece dell'incertezza non produce effetti apprezzabili entro la singola cifra significativa con la quale si esprime normalmente la barra di errore.

[1] Si può discutere a lungo sulla validità della somma in quadratura e sulle motivazioni (indipendenza delle due sorgenti di errore, carattere stocastico o non) che se ne possono

dare. Notate che, almeno quando le misure sono eseguite su portate adeguate al valore della grandezza misurata, la somma in quadratura porta in genere a risultati non significativamente diversi dalla somma tout-court.

Grafici, best-fit e Python

francesco.fuso@unipi.it

(Dated: version 5 - FF, 29 settembre 2018)

Questa nota intende richiamare alcuni metodi che sono normalmente coinvolti nell'analisi dei dati, cioè la preparazione di grafici e la realizzazione di best-fit. L'accento è posto soprattutto sugli aspetti pratici e a questo scopo si fa riferimento all'impiego del software Python (saccheggiando ampiamente i testi di Luca Baldini e Carmelo Sgrò, che conoscete, lots of kudos to them!). Lo scopo della nota è anche quello di chiarire l'ambito concettuale delle nostre analisi, cioè stabilire l'utilità del best-fit e gli scopi principali che esso avrà per noi.

I. GRAFICI

La rappresentazione di coppie di dati (sperimentali o no che siano) in un grafico è una tecnica diffusa e nota al punto che è certamente fuori luogo discuterne l'importanza. Spesso il grafico è il risultato principale, o addirittura unico, di un'attività scientifica. Esso deve pertanto contenere in modo ordinato e chiaro quante più informazioni possibile. Dunque le grandezze riportate sugli assi devono essere correttamente indicate assieme alla loro unità di misura [1], i valori numerici sugli assi devono essere ben leggibili [2], il range dei valori graficati deve essere opportuno (gli spazi vuoti su un grafico non servono a molto), e, almeno se i dati sono di origine sperimentale, devono *sempre* comparire le barre di errore, a prescindere dal fatto che esse vengano usate, o meno, in sede di analisi.

Un grafico decente deve poter mostrare al primo colpo se i dati seguono un qualche andamento. A tale scopo talvolta può essere utile la rappresentazione logaritmica o semi-logaritmica [3]. Per esempio, un andamento di tipo esponenziale decrescente (il decadimento temporale della carica sulle armature di un condensatore) è rappresentato in carta semi-logaritmica come un andamento lineare, con una pendenza inversamente proporzionale alla costante tempo di decadimento; un andamento secondo una legge di potenza è rappresentato in scala logaritmica come un andamento lineare, con una pendenza proporzionale all'esponente. Inoltre, a prescindere dall'esigenza di individuare al primo colpo un certo andamento, l'uso della rappresentazione logaritmica è opportuno ogni volta che si devono rappresentare dati che spaziano su un ampio range.

Anche se l'analisi comprende altre fasi, per esempio un best-fit come nell'esempio che tratteremo in seguito, è *fondamentale* che la visualizzazione del grafico *preceda* ogni altra operazione. Nello script di Python è quindi necessario che la parte relativa alla realizzazione del grafico *preceda* quella relativa al best-fit. Ciò consente di verificare immediatamente la presenza di eventuali errori nella "acquisizione" dei dati, manuale o automatizzata che sia, e di escludere errori nella parte di script che arriva alla preparazione del grafico stesso, che (quasi) sempre include la lettura dei dati da un file di testo.

A. Grafico in Python

Richiamiamo qui le principali operazioni (concettuali e pratiche) da compiere per realizzare un grafico in Python. Si fa riferimento ai pacchetti, o librerie, di uso comune per questi scopi, cioè `pylab` ed eventualmente `numpy`, da importare all'inizio dello script. Immaginiamo poi di avere un set di dati, corredata da incertezze, registrato in un file di testo presente nel computer (per vari ovvi motivi è fortemente *sconsigliato*, cioè *vietato*, scrivere direttamente l'array o lista di dati all'interno dello script!). In questo esempio, che conduce al grafico di Fig. 1, supponiamo che i dati siano stati inseriti in quattro colonne che riportano rispettivamente i valori x , Δx , y e Δy , con ovvio significato dei simboli.

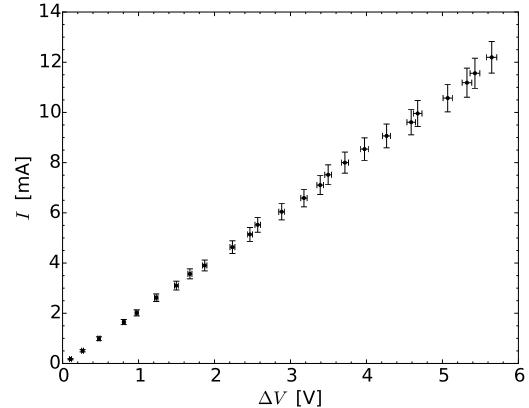


Figura 1. Grafico prodotto dallo script di Python discusso nel testo. Si notino: la presenza di barre di errore, la buona leggibilità dei caratteri, l'appropriata scelta del range rappresentato, la presenza delle unità di misura (in carattere non corsivo o "matematico").

La prima istruzione necessaria è quella che ordina al software di aprire e leggere il file, trasferendone il contenuto in quattro arrays, che chiamiamo `x`, `Dx`, `y`, `Dy`, con ovvio significato. Questo si può ottenere con il comando `x,Dx,y,Dy=pylab.loadtxt('filename.txt', unpack=True)`, dove il nome del file può contenere anche l'eventuale indirizzamento alla directory in cui esso si trova [4]. A questo punto la realizza-

zione del grafico con le barre di errore può essere eseguita usando il comando (notate l'ordine degli argomenti) `pylab.errorbar(x,y,dy,dx,linestyle = '', color = 'black', marker = '.')`, che contiene anche delle istruzioni molto ovvie che riguardano il formato (senza linea che congiunge i punti), il colore (nero) e il tipo di marker (un puntino, che in genere è la scelta preferibile per evitare di coprire le barre di errore). Per visualizzare il grafico sullo schermo, che al momento è tutto quello che serve, è sufficiente inserire il comando `pylab.show()`. Se il grafico deve essere salvato in oppor-

tuno formato, per esempio `.pdf`, per ulteriori usi o per la stampa [5], conviene usare i comandi a icona della finestra di rappresentazione grafica di Pylab. Prima dell'istruzione di visualizzazione vanno anteposte alcune istruzioni di stile, che comprendono anche la denominazione e l'unità di misura delle grandezze riportate sugli assi. Per fissare le idee, immaginiamo che i dati rappresentino delle intensità di corrente (I , misurate in milliAmpere, [mA]) in funzione di differenze di potenziale (ΔV , misurate in Volt, [V]).

Lo script di Python che produce il grafico di Fig. 1 è riportato qui nel seguente:

```
import pylab
# data load (nomefile can include address to the folder)
x,Dx,y,Dy=pylab.loadtxt('nomefile.txt',unpack=True)
# scatter plot with error bars
pylab.errorbar(x,y,Dy,Dx,linestyle = '', color = 'black', marker = '.')
# bellurie
pylab.rc('font',size=18)
# the $ symbol allows using LaTeX style characters (not so useful, though)
pylab.xlabel('$\Delta V$ [V]')
pylab.ylabel('$I$ [mA]')
pylab.title('Data plot')
pylab.minorticks_on()
# show the plot
pylab.show()
```

II. BEST-FIT

Eseguire un best-fit è una delle operazioni più frequenti sia nella pratica sperimentale che nelle attività “teoriche” (per esempio in simulazioni numeriche di esperimenti reali o concettuali). Il best-fit implica di analizzare quantitativamente le discrepanze tra dati e previsioni di un modello formulato sulla base di considerazioni fisiche. Il modello conduce generalmente a una funzione *analitica* $f(x)$, qui supposta dipendente dalla variabile unica x , munita di opportuni parametri p_m . Nel produrre il modello si è confidenti che la funzione possa descrivere le osservazioni sperimentali. Dunque il best-fit analizza le discrepanze tra i valori y_i e le “previsioni” $f(x_i)$, con x_i e y_i coppie di dati, e agisce sui parametri p_m della funzione $f(x)$ in modo da minimizzare per quanto possibile tali discrepanze.

Sulla base di queste premesse è ovvio che fare un best-fit *non* significa (solo, e in realtà affatto) far passare una curva analitica su un set di dati sperimentali, operazione in genere poco significativa. In particolare, l'obiettivo del best-fit non è (mai) quello di determinare una funzione $f(x)$ “qualsiasi” che sia in grado di descrivere al meglio le osservazioni. Infatti la costruzione della $f(x)$ deve essere fatta sulla base di considerazioni fisiche ben definite, con un grado di arbitrarietà idealmente molto limitato, o nullo. Sicuramente fare un best-fit conduce ad altri

importanti risultati, tra cui:

1. determinare quantitativamente grandezze incognite del sistema sotto analisi che compaiono come parametri, o entrano nei parametri, della funzione di fit e stimare l'incertezza a loro associata, permettendo, in sostanza, di eseguirne una “misura indiretta”;
2. qualche volta, se possibile, confrontare diversi modelli di interpretazione dei dati, ovvero diverse funzioni di fit, sempre create sulla base di considerazioni fisiche, per stabilire quale consenta la migliore descrizione dell'osservazione sperimentale;
3. solamente nei (rari, per noi) casi in cui le misure hanno incertezze di origine prevalentemente statistica, determinare la significatività dell'interpretazione.

Per gli scopi didattici che ci prefiggiamo, è *sempre* necessario specificare tra i *risultati del fit*:

1. l'espressione analitica della funzione modello, con chiara indicazione di quali sono i parametri liberi del best-fit;
2. il valore dei parametri della funzione ottenuti dal best-fit, che diventano grandezze fisiche, correttamente dimensionate;

3. l'incertezza su tali parametri, ovvero sulla misura indiretta delle grandezze che compaiono come parametri nella funzione di best-fit, assieme alla *modalità* con cui questa incertezza è stata determinata, secondo quanto illustreremo nel seguito di questa nota;
4. il valore del χ^2 risultante e del numero di gradi di libertà $ndof$, pari alla differenza tra numero di coppie di dati e numero di parametri del best-fit;
5. nel caso di fit a più di un parametro, l'esplicita indicazione della *covarianza normalizzata*, qualche volta detta anche correlazione tra i parametri, secondo quanto discuteremo nel seguito di questa nota.

In termini generali esistono diversi metodi per eseguire un best-fit. Qui useremo il cosiddetto metodo del *minimo χ^2* , che è un'estensione del fit a *minimi quadrati* adatta per trattare campioni di dati con incertezza non “uniforme”. Questo è quanto si verifica molto spesso nelle misure di segnali elettrici effettuate con i normali strumenti di laboratorio [6].

A. Richiami sul fit del minimo χ^2

Eseguire un best-fit dei *minimi quadrati* secondo la funzione $f(x)$ richiede di minimizzare la somma dei residui (quadrati), cioè di minimizzare la funzione

$$\sum_i [y_i - f(x_i)]^2 , \quad (1)$$

dove y_i e x_i sono le coppie di dati sperimentali e la somma è estesa al numero N di coppie di dati disponibili. Come ben sapete, questo metodo ha lo svantaggio di non considerare l'incertezza dei dati sperimentali Δy_i , che però può essere inserita come peso della somma nel seguente semplicissimo modo:

$$S = \sum_i \frac{[y_i - f(x_i)]^2}{(\Delta y_i)^2} ; \quad (2)$$

questa definizione permette di attribuire maggiore importanza nel best-fit a quei dati sperimentali che hanno l'incertezza minore, e viceversa minore importanza a quelli più incerti, procedura sicuramente sensata. Qua-lora tutte le incertezze Δy_i fossero uguali, il termine Δy_i al denominatore potrebbe essere messo in evidenza, e l'espressione da minimizzare diventerebbe analoga a quella di Eq. 1.

Alla grandezza S considerata in Eq. 2 si dà spesso (quasi sempre *impropriamente*) il nome di χ^2 . Questo nome nasce dall'*analoga* con la variabile aleatoria χ^2 costruita come somma dei quadrati di un'altra variabile aleatoria *standard* ξ_i : $\chi^2 = \sum_i \xi_i^2$. La distribuzione di probabilità del χ^2 è nota (tabelle e calcoli numerici) e si sa che essa, per un numero N sufficientemente grande, tende ad assumere il valore medio $\mu_{\chi^2} \simeq N$ e la deviazione standard

$\sigma_{\chi^2} \simeq \sqrt{2N}$. Questo si verifica solo se la ξ_i è distribuita secondo una Gaussiana (significato del termine standard) a media nulla e *varianza unitaria (normalizzata)*. Dunque nel considerare la grandezza S definita in Eq. 2 come un χ^2 stiamo facendo un'*importante e delicata affermazione*. Infatti riteniamo di poter sostituire la varianza σ_i^2 dell'(ipotetica) distribuzione delle ξ_i con $(\Delta y_i)^2$.

Per ora notiamo che *solo se* le condizioni che abbiamo posto sono ritenute ragionevoli il calcolo del χ^2 può essere usato per stabilire dei criteri per la valutazione *quantitativa* della significatività del best-fit. Il più noto di questi criteri è quello di Pearson, detto anche semplicemente del χ^2 , che avete conosciuto lo scorso anno. Esso è direttamente collegato alla stima della probabilità che si possa avere un χ^2 più alto, o più basso, di quello ottenuto, attraverso l'uso di tabelle che riportano l'integrale dell'area sottesa alla curva di distribuzione (normalizzata), cioè la probabilità (normalizzata). Nei (per noi rari) casi in cui può essere sensatamente impiegato, tale metodo risponde allora a uno dei “requisiti” che avevamo posto prima, quello di dare una valutazione quantitativa della significatività del best-fit.

In tutti gli altri casi *non* si può predire quale sia il valore che rende il χ^2 “ragionevole”, in particolare non si può affermare che $\chi_{rid}^2 = \chi^2/ndof$ dovrebbe avvicinarsi all'unità. Anzi, poiché la grandezza S dipende in modo inverso dalle Δy_i , è chiaro che una sovrastima delle incertezze, come spesso si verifica quando si hanno dati affetti da errori sistematici di calibrazione, può condurre a valori di S , ovvero del χ^2 , che possono idealmente tendere a zero.

In conclusione, la grandezza S , o per praticità χ^2 , rimane sempre rappresentativa della “qualità” del best-fit, e può per esempio essere usata per confrontare più modelli, essendo quello con il χ_{rid}^2 minore il modello migliore, ma non possono essere tratte conclusioni sulla significatività del best-fit a meno che le incertezze Δy_i non siano di carattere prevalentemente statistico.

B. Errore Δx_i

L'impiego di $(\Delta y_i)^2$ fatto in Eq. 2 implica evidentemente di poter trascurare l'incertezza Δx_i sulla misura della grandezza x_i . È chiaro che questa condizione non è sempre rispettata.

Il metodo del minimo χ^2 non contiene un'estensione immediata e rigorosa per considerare Δx_i . Esistono alcuni metodi numerici in cui la minimizzazione della grandezza di Eq. 2 è cercata muovendosi su opportune “traiettorie” dello spazio x, y , permettendo di tenere conto di tutte e due le incertezze. Questi metodi, implementabili anche con Python (metodo ODR), non hanno grande diffusione in ambito scientifico, poiché risultano normalmente affidabili solo per funzioni modello lineari, che si trovano piuttosto raramente. Dunque il loro impiego è fortemente sconsigliato, a meno che non sia supporta-

to dai risultati dell'approccio ordinario descritto qui nel seguito.

Questo approccio (non pulito e non sicuro) si basa sulla propagazione degli errori. In pratica si valuta il contributo sull'incertezza in y_i , $\Delta y_i|_{\Delta x_i}$, detto talvolta *errore equivalente*, dovuto all'incertezza su x_i , Δx_i . Secondo le regole della propagazione degli errori, tale contributo può essere espresso come $\Delta x_i |\partial f(x)/\partial x|_{x=x_i}$ per cui si può individuare una sorta di *errore efficace* sul dato i -esimo attraverso somma in quadratura:

$$(\Delta y_{i,eff})^2 = (\Delta y_i)^2 + \left(\Delta x_i \left| \frac{\partial f(x)}{\partial x} \right|_{x=x_i} \right)^2. \quad (3)$$

Questa espressione ha un'evidente limitazione, poiché richiede la conoscenza a priori della funzione $f(x)$ inclusi i suoi parametri, che invece è in genere proprio quanto si vuole determinare con la procedura di best-fit. Dunque la sua implementazione deve essere valutata attentamente caso per caso. Normalmente, è bene prima di tutto eseguire un best-fit trascurando le incertezze Δx_i ; i parametri ottenuti da questo best-fit possono poi essere usati per la valutazione di $\Delta y_i|_{\Delta x_i}$ consentendo la realizzazione di un nuovo best-fit (con una nuova stima dell'incertezza, come da Eq. 3). Qualora i parametri ottenuti dal nuovo best-fit si discostino in modo significativo dalla prima valutazione (in genere non succede mai), è possibile iterare la procedura.

III. BEST-FIT NUMERICO

In linea di principio è possibile minimizzare la grandezza S di Eq. 2 usando un approccio analitico, per esempio attraverso il metodo dei moltiplicatori di Lagrange. Tuttavia questa opzione è praticabile solo per funzioni semplicissime, per esempio quelle lineari, caso in cui si trovano le funzioni riportate in Appendice A. L'approccio numerico è di sicuro molto più “potente” e versatile, dato che può essere usato per funzioni virtualmente di ogni tipo. In questo approccio ci si deve fidare di procedure di calcolo eseguite da un computer (il cosiddetto “algoritmo”, in gergo). Nella quasi totalità dei software di trattamento dati, l'algoritmo è il Levenberg-Marquardt Algorithm (LMA), una procedura numerica per la ricerca di minimi locali generalmente efficiente e affidabile.

In Python la routine di minimizzazione è per esempio contenuta nel pacchetto `scipy.optimize`, da cui deve essere “estratta” e importata per renderla utilizzabile. I comandi relativi, come mostrato nel seguito, sono piuttosto semplici e di comprensione immediata. Prima di procedere nella descrizione dell'implementazione, occorre sottolineare un aspetto generale, critico, e potenzialmente laborioso, nell'uso dell'approccio numerico.

A prescindere dal tipo di funzione, è *sempre necessario fornire dei valori iniziali* alla routine per tutti i parametri contenuti nella funzione, da cui essa possa partire per

cercare il minimo. Soprattutto nel caso di funzioni “complicate” (bastano le trigonometriche), oppure quando è a priori possibile che ci sia più di un minimo locale nel χ^2 , è opportuno che le condizioni iniziali siano tali da fornire un “ragionevole” accordo con i dati da fittare. Se questo non si verifica, è possibile che l'algoritmo di minimizzazione non converga, oppure che converga a valori che, sulla base di considerazioni fisiche, sono paleamente sbagliati.

La scelta dei valori iniziali può essere fatta sulla base di diverse considerazioni. Per esempio, spesso la costruzione del modello permette di individuare a priori dei valori approssimati dei parametri della funzione di best-fit. Altre volte il modello stesso è così semplice (il caso lineare discusso in questa nota rappresenta bene questa condizione) che è possibile determinare valori ragionevoli dei parametri semplicemente ispezionando i dati. In genere, comunque, è sempre possibile “tentare” dei valori iniziali e vedere immediatamente dal grafico quant'è la distanza tra la funzione modello calcolata con tali valori iniziali e i dati.

Di conseguenza, è *estremamente opportuno* concepire l'approccio al best-fit come una successione di passi: i primi, che abbiamo già discusso, riguardano l'importazione dei dati e il loro grafico. A questi seguono quelli in cui si definisce la funzione modello, si impostano i parametri iniziali e si esegue il grafico dei dati sperimentali con sovrapposto quello della funzione *calcolata per i parametri iniziali*. Se organizzate il vostro lavoro, e anche lo script, secondo questa logica, vi troverete assai avvantaggiati quando dovrete eseguire dei best-fit un po' più complicati rispetto agli andamenti lineari trattati in questo esempio, riuscendo a terminare con la debita efficienza l'analisi dei dati.

Qui ci limitiamo ad applicare il best-fit numerico al set di dati già considerato in Fig. 1, di cui si sa che il modello prevede un andamento di tipo lineare conseguenza della legge di Ohm. Infatti la funzione che riteniamo possa descrivere i dati è $I_i = \Delta V_i/R + I_0$, che ha appunto la forma di una “retta che non passa per l'origine”, $f(x) = a + bx$, con $a = I_0$ e $b = 1/R$. Osservando il grafico si possono determinare in maniera pressoché immediata i valori iniziali dei parametri a , b (nel caso specifico, si può prendere per esempio $a_{in} = 0$ e $b_{in} = 2 \text{ ohm}^{-1}$).

L'istruzione di Python che lancia la routine di minimizzazione è `curve_fit`. Gli argomenti dell'istruzione sono, nell'ordine, il nome della funzione (che conviene definire nello stesso script), l'array da usare come variabile indipendente (nel nostro caso `x`), l'array dei dati dipendenti (nel nostro caso `y`), l'array dei valori iniziali dei parametri (che ha lunghezza pari al numero di parametri stesso, dunque 2 nel caso esaminato, e si chiama `init`), e l'array da usare come σ_i nel calcolo del χ^2 . Osservate che, qualora l'array delle σ_i non venga specificato, la routine considera un'incertezza costante unitaria, e, di fatto, il best-fit diventa dei minimi quadrati (non tornerà utile mai, o quasi mai, nel corso). Inoltre l'istruzione contiene un ulteriore importante comando che istruisce la routi-

ne sul “significato” da dare alle incertezze sperimentali: a questa istruzione si accede con l’opzione `absolute_sigma` ed essa, per la sua rilevanza, verrà trattata a parte nella sezione V.

Al termine della sua esecuzione, la routine restituisce nell’ordine questi due oggetti matematici: un array (nello script di esempio chiamato `pars`), che contiene il valore dei parametri ottenuto dal best-fit, e una matrice quadrata (`covm` nell’esempio), detta *matrice di covarianza*, il

cui significato sarà discusso in sezione IV. Sia array che matrice hanno dimensioni date dal numero di parametri della funzione: dunque nel nostro esempio l’array è un vettore di due elementi e la matrice comprende quattro elementi.

Un possibile script di Python per il best-fit numerico ai dati considerati (lo script si trova in rete con il nome `numerical_fit.py`) è riportato qui di seguito.

```

import pylab
import numpy
from scipy.optimize import curve_fit

# data load
x,Dx,y,Dy=pylab.loadtxt('nomefile.txt',unpack=True)
# scatter plot with error bars
pylab.errorbar(x,y,Dy,Dx,linestyle = '', color = 'black', marker = '.')

# bellurie
pylab.rc('font',size=18)
pylab.xlabel('$\Delta V$ [V]')
pylab.ylabel('I [mA]')
pylab.minorticks_on()

# AT THE FIRST ATTEMPT COMMENT FROM HERE TO THE END

# define the function (linear, in this example)
def ff(x, aa, bb):
    return aa+bb*x

# define the initial values (STRICTLY NEEDED!!!)
init=(0,2)

# prepare a dummy xx array (with 2000 linearly spaced points)
xx=numpy.linspace(min(x),max(x),2000)

# plot the fitting curve computed with initial values
# AT THE SECOND ATTEMPT THE FOLLOWING LINE MUST BE COMMENTED
pylab.plot(xx,ff(xx,*init), color='blue')

# set the error
sigma=Dy
w=1/sigma**2

# call the minimization routine
pars,covm=curve_fit(ff,x,y,init,sigma, absolute_sigma=False)
# calculate the chisquare for the best-fit function
chi2 = ((w*(y-ff(x,*pars))**2)).sum()
# determine the ndof
ndof=len(x)-len(init)

# print results on the console
print('pars:',pars)
print('covm:',covm)
print ('chi2, ndof:',chi2, ndof)

```

```
# plot the best fit curve
pylab.plot(xx,ff(xx,*pars), color='red')

# show the plot
pylab.show()
```

Osservate che uno script concepito in questo modo rende estremamente semplice eseguire controlli preliminari sulla varie fasi elencate in precedenza (grafico dei dati e grafico con sovrapposta la curva di best-fit calcolata per i parametri iniziali). A questo scopo è sufficiente commentare determinate parti dello script: si ricorda che per commentare una singola linea basta anteporre il carattere `#`, mentre per commentare parti più estese può essere conveniente racchiuderle tra linee contenenti un triplo apice (`'''`).

Alcune osservazioni potenzialmente rilevanti sullo script: come potete notare, il grafico della funzione è eseguito usando un array dummy per la variabile indipendente, in questo esempio composto da 2000 punti equispaziati tra minimo e massimo delle `x`, e non i valori (sperimentali) dell'array `x`. Il motivo è ovvio: nel caso in cui il numero di coppie di dati acquisiti fosse piccolo, la rappresentazione della funzione somiglierebbe a una spezzata. Questo non sarebbe corretto non solo per motivi "estetici", ma anche perché una spezzata lascia il dubbio su cosa succede tra un punto e quello successivo. Notate anche la semplicità con cui la funzione viene valutata con diversi parametri: per intenderci, per graficare la funzione valutata sui parametri iniziali è sufficiente scrivere `pylab.plot(xx,ff(xx,*init))`, per graficare quella valutata con i parametri ottenuti al termine della minimizzazione `pylab.plot(xx,ff(xx,*pars))`. Infine notate la sintassi, tipica di `numpy`, con cui viene eseguita la somma degli elementi dell'array per il calcolo del χ^2 e ricordate che non c'è alcun bisogno di "formattare" in alcun modo i risultati che compaiono sulla console (ormai siete grandi!).

L'esito del best-fit è mostrato, sovrapposto ai dati sperimentali, in Fig. 2. I risultati del fit, includendo anche la covarianza normalizzata, o correlazione, e l'indicazione dell'opzione di calcolo delle incertezze sui parametri, di cui discuteremo in seguito, sono:

$$I_0 = (-48 \pm 8) \text{ mA} \quad (4)$$

$$R = (470 \pm 4) \text{ ohm} \quad (5)$$

$$\text{norm.cov.} = -0.51 \quad (6)$$

$$\chi^2/\text{ndof} = 1.4/23 \quad (7)$$

$$\text{absolute_sigma} = \text{False}. \quad (8)$$

Osservate come i risultati del best-fit siano stati messi in modo che le grandezze fisiche di interesse, I_0 e R , venissero tutte e due esplicitate (la propagazione dell'errore è stata ovviamente usata per specificare l'incertezza di R).

Il best-fit del minimo χ^2 qui condotto è stato eseguito trascurando l'incertezza Δx_i sulle grandezze x_i . Se e

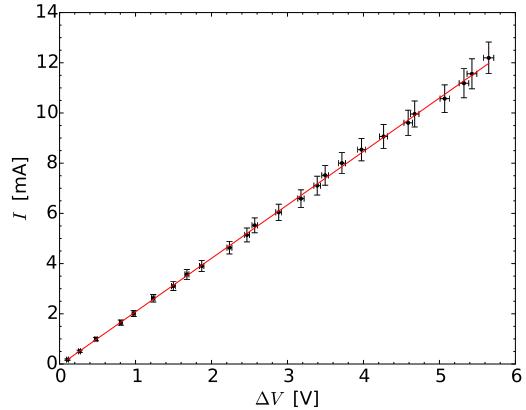


Figura 2. Grafico dei dati mostrati in Fig. 1 con sovrapposto il best-fit ottenuto per via numerica (linea continua rossa).

quanto questa scelta sia ragionevole dipende, ovviamente, dai dati acquisiti. Per esercizio, possiamo provare a considerare questa incertezza nel best-fit secondo la procedura stabilita in sezione II B. Nel caso lineare che stiamo considerando, l'espressione da impiegare per costruire l'incertezza efficace è semplicissima. Si ottiene infatti $(\Delta y_{i,\text{eff}})^2 = (\Delta y_i)^2 + (b\Delta x_i)^2$.

L'implementazione pratica è anche semplicissima: nella riga di script in cui si "definiscono le incertezze" (il commento è `# set the error`) è sufficiente sostituire la riga `sigma = Dy` con `sigma = numpy.sqrt(Dy**2+(bbb*Dx)**2)`, dove `bbb` rappresenta il valore del parametro b del fit. Come già discusso, questo parametro può essere ottenuto eseguendo, prima, un best-fit in cui l'incertezza dei dati è rappresentata solo da Δy_i .

Nel caso in esempio considerare l'errore efficace non conduce a modifiche significative dei risultati espressi in Eqq. 4-8, a parte, come atteso, una ulteriore diminuzione del χ^2 (che diventa $\chi^2 = 1.3$). Se ne conclude che, per l'esempio considerato, le incertezze Δx_i giocano un ruolo effettivamente trascurabile.

IV. LA MATRICE DI COVARIANZA

L'uscita dell'algoritmo di minimizzazione comprende una matrice C (la `covm` dello script) detta di covarianza. Questa matrice è simmetrica e gli elementi sulla diagonale rappresentano il quadrato delle incertezze che il software attribuisce ai parametri del best-fit, cioè $\sqrt{C_{mm}} = \Delta p_m$.

Gli elementi fuori diagonale, $C_{ij} = C_{ji}$, sono invece rappresentativi della cosiddetta *covarianza* tra i parametri. Questi valori danno una misura di quanto la variazione di un parametro influenzi la variazione dell'altro.

Nel caso, semplice, di una retta che non passa per l'origine, come quello che stiamo considerando, è facilissimo rendersi conto che le variazioni della pendenza (parametro b) e dell'intercetta (parametro a , si intende intercetta con l'asse verticale del grafico) sono legate fra loro. In altre parole, la variazione di un parametro può essere “compensata” dalla variazione dell'altro. Se, come in questo esempio, l'aumento di un parametro è compensato dalla diminuzione dell'altro, la covarianza è negativa, altrimenti essa è positiva. Una covarianza nulla, un caso praticamente irrealizzabile nei best-fit di interesse fisico, indica che i parametri sono completamente indipendenti tra loro, cioè che la variazione dell'uno non è affatto correlata con quella dell'altro.

Per quantificare la covarianza si crea una quantità, detta *covarianza normalizzata*, o *coefficiente di correlazione*, che varia tra -1 e 1. Essa è data da

$$c_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} . \quad (9)$$

Conventionalmente, i parametri si dicono fortemente correlati (o anticonrelati, in caso di segno negativo) se $|c_{ij}| \gtrsim 0.8$ e totalmente correlati (o anticonrelati) se $|c_{ij}| \approx 1$ [7]. Nell'esempio considerato in questa nota i due parametri di fit sono scarsamente (anti)correlati fra loro. Parametri totalmente correlati o quasi completamente scorrelati sono sospetti, e possono indurre a chiedersi se il modello impiegato è quello realmente più adatto a descrivere le osservazioni sperimentali.

La covarianza, nella sua forma normalizzata, è una grandezza che, almeno per gli scopi didattici del nostro corso, deve essere *sempre* inclusa nei risultati di ogni best-fit a più di un parametro, a meno di diverse richieste, come informazione di controllo sulla “qualità” del best-fit.

Inoltre la covarianza (non normalizzata) è certamente utile quando i risultati di un best-fit vengono impiegati per fare delle “previsioni”. Per esempio, nel caso che stiamo considerando potrebbe essere richiesto di determinare l'intensità di corrente I' prevista per un certo valore $\Delta V'$ della differenza di potenziale: una previsione accurata richiede di tenere in debito conto della covarianza.

A. Richiami sulla covarianza

Immaginiamo di avere una funzione $f(\alpha, \beta)$ che dipende da due variabili α, β e supponiamo che queste variabili corrispondano a due grandezze misurabili. La loro misura fornisce i valori α_i, β_i , con i che corre da 1 a N numero totale delle misure. In seguito alle misure, si possono determinare i valori medi $\bar{\alpha}$ e $\bar{\beta}$ e le rispettive varianze (sperimentali), definite come

$$\sigma_\alpha^2 = \frac{1}{N-1} \sum_i (\alpha_i - \bar{\alpha})^2 \quad (10)$$

$$\sigma_\beta^2 = \frac{1}{N-1} \sum_i (\beta_i - \bar{\beta})^2 . \quad (11)$$

La varianza della grandezza $f_i = f(\alpha_i, \beta_i)$ sarà definita come

$$\sigma_f^2 = \frac{1}{N-1} \sum_i [f(\alpha_i, \beta_i) - \bar{f}]^2 , \quad (12)$$

con $\bar{f} = f(\bar{\alpha}, \bar{\beta})$. Sviluppiamo al primo ordine di Taylor la funzione $f(\alpha_i, \beta_i)$ attorno a \bar{f} :

$$f(\alpha_i, \beta_i) \simeq \bar{f} + (\alpha_i - \bar{\alpha}) \frac{\partial f}{\partial \alpha} + (\beta_i - \bar{\beta}) \frac{\partial f}{\partial \beta} . \quad (13)$$

Introducendo lo sviluppo nell'Eq. 12, si ottiene

$$\sigma_f^2 \simeq \frac{1}{N-1} \sum_i \left[(\alpha_i - \bar{\alpha}) \frac{\partial f}{\partial \alpha} + (\beta_i - \bar{\beta}) \frac{\partial f}{\partial \beta} \right]^2 = (14)$$

$$= \sigma_\alpha^2 \left(\frac{\partial f}{\partial \alpha} \right)^2 + \sigma_\beta^2 \left(\frac{\partial f}{\partial \beta} \right)^2 + 2\sigma_{\alpha\beta} \frac{\partial f}{\partial \alpha} \frac{\partial f}{\partial \beta} , \quad (15)$$

dove si definisce *covarianza* (sperimentale) delle grandezze α_i e β_i [8] la

$$\sigma_{\alpha\beta} \equiv \frac{1}{N-1} \sum_i (\alpha_i - \bar{\alpha})(\beta_i - \bar{\beta}) . \quad (16)$$

Naturalmente l'espressione di Eq. 14 può essere facilmente generalizzata al caso in cui i parametri siano più di due: è sufficiente sommare gli ulteriori termini con le varianze e gli ulteriori “doppi prodotti”.

Per semplificare ulteriormente la trattazione e per avvicinarci al caso di nostro interesse, supponiamo $f(\alpha, \beta) = (\alpha + \beta)$. In questo caso si ottiene immediatamente

$$\sigma_f^2 \simeq \sigma_\alpha^2 + \sigma_\beta^2 + 2\sigma_{\alpha\beta} . \quad (17)$$

La funzione di best-fit lineare $f = a + bx$ può essere interpretata come somma di due termini, cioè possiamo porre $\alpha = a$ e $\beta = bx$. Allora la varianza (sperimentale) sulla funzione può essere espressa dall'Eq. 17. Possiamo poi interpretare la procedura di best-fit come una sorta di esperimento (l'esperimento consiste nel disegnare una retta che passa per le barre di errore dei dati sperimentali) e la matrice di covarianza come indicativa delle varianze associate a questo esperimento. Questo equivale a porre $\sigma_\alpha^2 = C_{11}$, $\sigma_\beta^2 = C_{22}x^2$, $\sigma_{\alpha\beta} = C_{12}x$ (la presenza di x^2 e x è dovuta alla definizione di $\beta = bx$ ed è necessaria dal punto di vista dimensionale).

Fatte queste premesse, l'utilità pratica della covarianza è facile da capire. Avendo eseguito il best-fit dei nostri dati e ottenuto il risultato di Eqq. 4-8, supponiamo di voler predire il valore I' dell'intensità di corrente che corrisponde a una certa differenza di potenziale $\Delta V'$. Sapendo che l'andamento è lineare secondo la legge $I = \Delta V/R + I_0$, ovvero $y = a + bx$, con a e b determinati dal fit assieme alle loro incertezze Δa e Δb , avremo $I' = \Delta V'/R + I_0$, ovvero $y' = a + bx'$, con ovvio significato dei termini.

Per determinare l'incertezza $\Delta I'$ da attribuire al valore I' , ovvero $\Delta y'$ da attribuire a y' previsto, potremo utilizzare l'Eq. 17:

$$\Delta y' = \sigma_f = \sqrt{C_{11} + C_{22}x'^2 + 2C_{12}x'} \quad (18)$$

ovvero

$$\Delta I' = \sqrt{(\Delta a)^2 + (\Delta b)^2 \Delta V'^2 + 2c_{12}\Delta a \Delta b \Delta V'}, \quad (19)$$

dove abbiamo usato la covarianza normalizzata definita in Eq. 9.

Se per esempio supponiamo $\Delta V' = 1$ V (senza incertezza), usando il risultato di Eqq. 4-8 troviamo $I' = (50 \pm 8)$ mA. Notate che il segno negativo della covarianza “fa diminuire” l'incertezza sulla previsione rispetto a quanto si avrebbe applicando la propagazione dell'errore a dati indipendenti tra loro [9]. Questa circostanza può essere qualitativamente generalizzata. Supponiamo di avere dei dati sperimentali che possono essere fittati con diverse funzioni modello, contenenti un diverso numero di parametri. Per esempio, immaginiamo che i dati possano essere rappresentati da una retta che passa o non passa per l'origine, cioè con funzioni che hanno rispettivamente uno e due parametri liberi (fisicamente, la situazione potrebbe essere quella di un andamento lineare in cui c'è un piccolo offset costante, che viene ritenuto trascurabile a priori se nel fit si usa la retta che passa per l'origine). Nel caso di covarianza negativa, il parametro con lo stesso significato, cioè la pendenza della retta, uscirà dal fit con un'incertezza minore se il fit stesso è eseguito con due parametri liberi. Viceversa, se la covarianza è positiva, esso uscirà presumibilmente con un'incertezza maggiore.

V. ABSOLUTE SIGMA

Il titolo criptico di questa sezione è legato a una particolarità della procedura di best-fit di Python [10]: tra le opzioni del comando `curve_fit` ce ne è una che suona piuttosto misteriosa, e rimane tale anche dopo aver consultato la scarna e confusa documentazione disponibile in rete.

L'opzione in questione è `absolute_sigma`, che può avere come valore `False` (*di default*, cioè attiva anche senza porre alcuna specifica nella linea di comando di `curve_fit`) oppure `True` [11]. Potete verificare facilmente come la scelta di questa opzione non modifichi il valore dei parametri risultanti dal best-fit, ma come essa agisca, spesso in maniera non trascurabile, sulla determinazione dell'*incertezza* da attribuire al valore dei parametri.

Nel nostro percorso di “spiegazione” per l'esistenza di questa opzione, partiamo da un'ovvia considerazione: la maggior parte dei metodi di data reduction and analysis discendono dalla circostanza di avere a disposizione campioni di dati in cui l'errore ha un'origine prevalentemente statistica. Per esempio la propagazione dell'errore e il

metodo del minimo χ^2 presuppongono di trattare grandezze affette (principalmente) da errore statistico. Come già abbiamo sottolineato, la realtà sperimentale tipica delle misure di segnali elettrici è, purtroppo, ben diversa, a causa della pressoché inevitabile preponderanza dell'errore sistematico dovuto alla calibrazione. Di conseguenza, le condizioni nelle quali ci troviamo ad operare, per necessità, o, qualche volta, per praticità, sono le peggiori per l'applicazione rigorosa dei metodi di data reduction and analysis [12].

È facilissimo verificare cosa determina la scelta `absolute_sigma = True` rispetto a quella, di default (dunque valida anche senza specificare l'opzione) `absolute_sigma = False`. Detta Δp_m l'incertezza sul parametro p_m del best-fit (per intenderci, nel caso della retta e usando la nomenclatura precedente si ha $p_1 = a$ e $p_2 = b$), si ha

$$\Delta p_m|_{\text{False}} = \sqrt{\chi_{rid}^2} \Delta p_m|_{\text{True}}. \quad (20)$$

In altre parole, l'incertezza sui parametri di fit data dall'opzione di default (`False`) è “pesata” con la radice quadrata del χ_{rid}^2 .

Come illustreremo nel seguito, l'opzione `True` conduce a una valutazione dell'errore sui parametri che è in linea con le aspettative naïf. Supponiamo infatti di voler fittare dai dati a una retta: operando in modo grafico (con matita e righello), possiamo facilmente dedurre che maggiore è l'entità delle barre di errore dei nostri dati, maggiore è l'incertezza con cui possiamo determinare i parametri della retta di fit, cioè intercetta e coefficiente angolare. Questo è in effetti quanto si verifica con la procedura `analitica` di best-fit lineare riportata in Appendice A, dove, in sostanza, l'incertezza sui parametri è calcolata usando le regoline della propagazione dell'errore.

Partendo dalla supposizione che chi ha scritto la procedura `curve_fit` volesse, con la scelta di default `False`, soddisfare le esigenze di best-fit più comuni [13], che effettivamente sono quelle in cui l'errore sistematico prevale, possiamo cercare un rationale per la normalizzazione effettuata dalla procedura stessa. È ovvio che, nel caso di sovrastima delle barre di errore, come si ha quando l'errore sistematico prevale (almeno nelle nostre esperienze tipiche), il χ_{rid}^2 tende a valere meno di uno, per cui la normalizzazione di Eq. 20 agisce in modo da ridurre l'incertezza sulla valutazione dei parametri. Viceversa, nel caso in cui le barre di errore fossero, per qualche motivo, sottostimate, il χ_{rid}^2 tenderebbe a essere ben maggiore di uno, e quindi la normalizzazione agirebbe in senso opposto. Per un best-fit “a regola d'arte”, cioè realizzato con una appropriata funzione modello e facendo uso di incertezze di origine rigorosamente stocastica, $\chi_{rid}^2 \simeq 1$ e la normalizzazione non avrebbe alcun effetto sull'incertezza dei parametri, che risulterebbe la stessa a prescindere dall'opzione scelta.

A. Asymptotic vs standard

La definizione comunemente accettata per l'incertezza sui parametri risultanti da un best-fit è la seguente [14]: Δp_m costituisce l'intervallo di variazione di p_m che corrisponde a un aumento unitario del χ^2 . Questa definizione è in linea con quelle di incertezza per altre grandezze, misurate o calcolate, per cui essa prende qualche volta il nome di *standard error*.

Accanto a questa definizione ne esiste, anche se in forma un po' clandestina, un'altra, a cui qualche volta si fa riferimento con il nome *asymptotic error*. Nel caso di misure, il carattere asintotico significa che questa incertezza sarebbe quella determinata su un campione contenente un numero infinito di elementi. Se l'errore è dominato da fluttuazioni statistiche, allora le due definizioni di incertezza coincidono.

Nel caso del best-fit su dati pesati, cioè dotati di barre di errore, l'errore asintotico sui parametri del best-fit può essere inteso come l'incertezza che si avrebbe se il numero di dati fosse molto grande e se la loro incertezza di misura fosse di origine puramente stocastica. In queste condizioni il metodo del minimo χ^2 stabilisce $\chi^2_{rid} \simeq 1$ (il best-fit “a regola d'arte”, secondo quanto scritto prima). Allora possiamo rifrasare la definizione di errore asintotico sui parametri del best-fit, dicendo che esso rappresenta l'incertezza che sarebbe attribuita alla valutazione dei parametri *se fosse* $\chi^2_{rid} \simeq 1$.

In sostanza, la procedura `curve_fit` di Python consente di scegliere le due definizioni di errore sui parametri:

```
absolute_sigma = False → asymptotic error
absolute_sigma = True → standard error .
```

Dimostrare che la normalizzazione attraverso $\sqrt{\chi^2_{rid}}$ espressa in Eq. 20 conduce effettivamente a quanto affermato sopra è compito matematicamente complicato. Qualche hint può essere trovato in [15] e, meglio ancora, in [16], dove si illustra un pacchetto di Python, chiamato `kmpfit`, che, basandosi su `scipy.optimize`, permette di eseguire best-fit con un certo numero di funzionalità aggiuntive pre-assemblate.

Dal punto di vista operativo, le osservazioni che abbiamo svolto suggeriscono che:

1. se gli errori dei dati sperimentali sono di origine prevalentemente stocastica, l'opzione da usare è `True`;
2. se invece l'origine degli errori è poco conosciuta, oppure se si sa che il contributo sistematico è dominante, è *consigliata* l'opzione `False`, quella di default, che consente di evitare marchiane sovrastime (in qualche caso sottostime) dell'incertezza sui valori di parametri di best-fit.

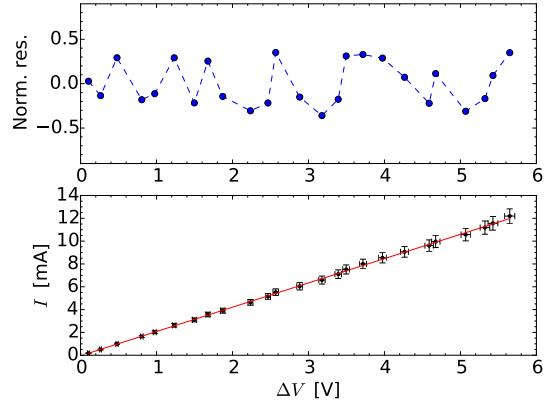


Figura 3. Grafici dei dati con sovrapposto il best-fit numerico (in basso) e dei residui normalizzati (in alto).

VI. RESIDUI NORMALIZZATI

Un'ulteriore analisi (qualitativa) che è spesso utile compiere a posteriori sul best-fit è quella detta *dei residui*. I residui sono, come noto, le differenze $y_i - f(x_i)$, dove la funzione $f(x_i)$ è calcolata con i parametri ottenuti dal best fit. Nel caso di best-fit del minimo χ^2 possono essere analizzati in particolare i *residui normalizzati* r_i definiti come

$$r_i = \frac{y_i - f(x_i)}{\Delta y_i}, \quad (21)$$

nel caso in cui siano trascurabili le incertezze Δx_i (questo è il caso a cui faremo riferimento qui).

Idealmente, cioè per un campione di dati corredata da incertezze di origine statistica, per cui $\Delta y_i = \sigma_i$, e ben descritto dalla funzione modello prescelta, r_i dovrebbe essere una variabile aleatoria standard, con una distribuzione Gaussiana a media nulla e varianza unitaria. Dunque un'analisi corretta dei residui (normalizzati) potrebbe essere compiuta costruendo l'istogramma della distribuzione di r_i e valutando se esso è ben descritto da una Gaussiana. Tuttavia questa operazione ha senso solo in presenza di un campione di tante misure (N molto grande), che non è la situazione del nostro esempio.

Tuttavia si può sempre costruire un grafico dei residui (normalizzati) r_i in funzione di x_i e valutare “a occhio” se questo grafico presenta degli andamenti evidenti. Può infatti verificarsi che i residui siano distribuiti in modo paleamente disomogeneo. Supponendo una funzione di fit lineare, come in questa nota, se i dati contengono delle non linearità queste possono essere messe in evidenza dall'analisi dei residui. Infatti le discrepanze tra dati e previsioni del fit (lineare) sono attese essere più marcate agli estremi dell'intervallo di dati considerato. Inoltre nella pratica sperimentale succede spesso che i dati vengano acquisiti in modo “sequenziale”, uno dopo l'altro. Esistono dei *rumori* periodici, cioè, nel linguaggio di

questa nota, degli errori sistematici (definiremo qualitativamente il “rumore” in altra sede), che possono sovrapporsi all’esito della misura. Il carattere sistematico può talvolta essere evidenziato da un andamento ciclico nel grafico dei residui.

La Fig. 3 riporta un grafico analogo a quello di Fig. 2 assieme al grafico dei residui normalizzati. Si vede come,

nel caso considerato, non ci siano particolari andamenti, per cui nulla si può concludere da questa analisi. La figura è stata costruita con lo script disponibile in rete con il nome `numerical_fit_res.py`, che propone anche l’impiego dei subplots, un modo per rappresentare su una stessa figura diversi grafici. Esso è riportato qui di seguito.

```

import pylab
import numpy
from scipy.optimize import curve_fit
# data load
x,Dx,y,Dy=pylab.loadtxt('nomefile.txt',unpack=True)
# use subplots to display two plots in one figure
# note the syntax
pylab.subplot(2,1,2)
# scatter plot with error bars
pylab.errorbar(x,y,Dy,Dx,linestyle = '', color = 'black', marker = '.')
# bellurie
pylab.rc('font',size=18)
pylab.xlabel('$\Delta V$ [V]')
pylab.ylabel('$I$ [mA]')
pylab.minorticks_on()
# AT THE FIRST ATTEMPT COMMENT FROM HERE TO THE END
# define the function (linear, in this example)
def ff(x, aa, bb):
    return aa+bb*x
# define the initial values (STRICTLY NEEDED!!!)
init=(0,2)
# prepare a dummy xx array (with 2000 linearly spaced points)
xx=numpy.linspace(min(x),max(x),2000)
# plot the fitting curve computed with initial values
# AT THE SECOND ATTEMPT THE FOLLOWING LINE MUST BE COMMENTED
pylab.plot(xx,ff(xx,*init), color='blue')
# set the error
sigma=Dy
w=1/sigma**2
# call the minimization routine
pars,covm=curve_fit(ff,x,y,init,sigma, absolute_sigma=False)
# calculate the chisquare for the best-fit function
chi2 = ((w*(y-ff(x,*pars))**2)).sum()
# determine the ndof
ndof=len(x)-len(init)
# print results on the console
print('pars:',pars)
print('covm:',covm)
print ('chi2, ndof:',chi2, ndof)
# plot the best fit curve
pylab.plot(xx,ff(xx,*pars), color='red')

# switch to the residual plot
pylab.subplot(2,1,1)
# build the array of the normalized residuals
r = (y-ff(x,*pars))/sigma
# bellurie
pylab.rc('font',size=18)
pylab.ylabel('Norm. res.')

```

```

pylab.minorticks_on()
# set the vertical range for the norm res
pylab.ylim((- .9, .9))
# plot residuals as a scatter plot with connecting dashed lines
pylab.plot(x,r,linestyle="--",color='blue',marker='o')

# show the plot
pylab.show()

```

APPENDICE A: BEST-FIT LINEARE ANALITICO

La minimizzazione dell'Eq. 2 rispetto ai parametri a , b nel caso di $f(x) = a + bx$ può essere eseguita usando il metodo dei moltiplicatori di Lagrange. Il calcolo, che è abbastanza laborioso (in Appendice B se ne riporta una versione semplificata), conduce alle seguenti formule per i parametri a e b con le rispettive incertezze Δa e Δb [8]:

$$w_i = \frac{1}{\sigma_i^2} \quad (22)$$

$$\Delta' = \sum_i w_i \sum_i w_i x_i^2 - (\sum_i w_i x_i)^2 \quad (23)$$

$$a = \frac{\sum_i w_i x_i^2 \sum_i w_i y_i - \sum_i w_i x_i \sum_i w_i x_i y_i}{\Delta'} \quad (24)$$

$$\Delta a = \sqrt{\frac{\sum_i w_i x_i^2}{\Delta'}} \quad (25)$$

$$b = \frac{\sum_i w_i \sum_i w_i x_i y_i - \sum_i w_i x_i \sum_i w_i y_i}{\Delta'} \quad (26)$$

$$\Delta b = \sqrt{\frac{\sum_i w_i}{\Delta'}}, \quad (27)$$

dove le somme si intendono estese su tutte le N coppie di dati disponibili.

Le Eqq. 22-27 hanno un aspetto sicuramente poco simpatico, ma, con un minimo di attenzione, si può notare che in esse compaiono "solo" cinque espressioni indipendenti, tutte generate da somme di elementi (dati sperimentali x_i , y_i e pesi statistici $w_i = 1/\sigma_i^2$). Uno script di Python può assai facilmente calcolare tutto ciò. Lo script, che è disponibile in rete sotto il nome di `anal_lin_two_parms.py`, può per esempio avere la seguente forma:

```

import pylab
import numpy
# data load
x,Dx,y,Dy=pylab.loadtxt('nomefile.txt',unpack=True)
# scatter plot with error bars
pylab.errorbar(x,y,Dy,Dx,linestyle = '', color = 'black', marker = 'o')
# bellurie
pylab.rc('font',size=16)
pylab.xlabel('$\Delta V$ [V]')
pylab.ylabel('$I$ [mA]')
pylab.title('Data plot w analytical fit')
pylab.minorticks_on()
# set the error and the statistical weight
sigma=Dy
w=1/sigma**2
# determine the coefficients
c1=(w*x*x).sum(); c2=(w*y).sum();c3=(w*x).sum()
c4=(w*x*y).sum(); c5=(w).sum()
Dprime=c5*c1-c3**2
a=(c1*c2-c3*c4)/Dprime
b=(c5*c4-c3*c2)/Dprime
Da=numpy.sqrt(c1/Dprime)
Db=numpy.sqrt(c5/Dprime)
# define the linear function
# note how parameters are entered
# note the syntax
def ff(x, aa, bb):
    return aa+bb*x

```

```

# calculate the chisquare for the best-fit function
chi2 = ((w*(y-ff(x,a,b))**2)).sum()
# determine the ndof
ndof=len(x)-2
# print results on the console
print(a,Da, b,Db)
print (chi2, ndof)
# prepare a dummy xx array (with 2000 linearly spaced points)
xx=numpy.linspace(min(x),max(x),2000)
# plot the fitting curve
pylab.plot(xx,ff(xx,a,b), color='red')
# show the plot
pylab.show()

```

Il grafico che si ottiene applicando il best-fit analitico ai dati considerati in questa nota è indistinguibile da quello di Fig. 2 e i risultati sono:

$$I_0 = (-48 \pm 24) \text{ mA} \quad (28)$$

$$R = (470 \pm 6) \text{ ohm} \quad (29)$$

$$\chi^2/\text{ndof} = 1.4/23, \quad (30)$$

dove, rispetto a quelli di Eq. 4, si osserva un diverso valore delle incertezze sui parametri del best-fit, dovuta alla circostanza, già discussa, che nel fit numerico si era usata l'opzione `absolute_sigma = False`.

Per rendere completo il set di risultati del best-fit (a più di un parametro) manca la covarianza normalizzata, che qui non determiniamo perché il suo calcolo in forma analitica è piuttosto noioso.

Le formule di Eqq. 22-27 possono essere considerate come un'estensione del caso, matematicamente più semplice, di una retta che passa per l'origine [cioè una funzione del tipo $f(x) = bx$, che esprime un andamento direttamente proporzionale tra y e x]. La minimizzazione analitica del χ^2 corrispondente può essere eseguita in maniera molto semplice, grazie alla presenza di un unico parametro. In alternativa, si può lavorare sulle Eqq. 22-27 "imponendo" $a = 0$. Alla fine si ottiene

$$w_i = \frac{1}{\sigma_i^2} \quad (31)$$

$$\Delta' = \Sigma_i w_i \Sigma_i w_i x_i^2 - (\Sigma_i w_i x_i)^2 \quad (32)$$

$$b = \frac{\Sigma_i w_i x_i y_i}{\Sigma_i w_i x_i^2} \quad (33)$$

$$\Delta b = \sqrt{\frac{\Sigma_i w_i}{\Delta'}}, \quad (34)$$

cioè un set di equazioni un po' più maneggevole che nel caso precedente.

Se, invece, si impone $b = 0$, allora il calcolo si riconduce a quello della *media pesata*. Il risultato, come noto, è

$$w_i = \frac{1}{\sigma_i^2} \quad (35)$$

$$a = \frac{\Sigma_i w_i y_i}{\Sigma_i w_i} \quad (36)$$

$$\Delta a = \frac{\sigma}{\sqrt{N}}, \quad (37)$$

con σ la deviazione standard di una delle misure che compongono la media pesata.

APPENDICE B: FORMULE ANALITICHE PER IL BEST-FIT A UNA RETTA

Questa Appendice è dedicata a una dimostrazione (semplificata, per quanto possibile) delle Eqq. 22-27. Partiamo da una situazione generale, in cui supponiamo di avere una funzione modello generica $f(x)$ che contiene diversi parametri p_m (quelli da tenere liberi nel best-fit). Il problema del best-fit si riduce a determinare il set di parametri p_m che minimizzano la Eq. 2, affermazione che porta a determinare il set di parametri per cui

$$\frac{\partial \left[\sum_i \frac{(y_i - f(x_i))^2}{(\Delta y_i)^2} \right]}{\partial p_m} = 0, \quad (38)$$

dove si intende che viene implementato in qualche modo il controllo che la derivata nulla non corrisponda a un massimo.

Poiché la funzione $f(x)$ dipende dall'*intero* set di parametri, il problema è un buon esempio di ricerca di *minimo condizionato*. In termini generali, il minimo condizionato si può trovare con il metodo dei *moltiplicatori di Lagrange*, che probabilmente avete studiato o studierete in altra sede. Questo metodo conduce facilmente a una matematica abbastanza ostica, incompatibile con gli scopi di questa Appendice. Fortunatamente, se ci si restringe ad esaminare il caso di interesse per questa nota, la retta $f(x) = a + bx$ che ha parametri $p_1 = a$ e $p_2 = b$, la soluzione può essere determinata in maniera sufficientemente rapida anche senza fare uso del metodo generale dei moltiplicatori. Inoltre, al puro scopo di ripulire ulteriormente le formule, possiamo immaginare di partire con un fit dei minimi quadrati, in cui, in sostanza, supponiamo $\Delta y_i = 1$ in Eq. 38 (generalizzeremo poi al fit del minimo χ^2 che è di nostro interesse).

Grazie alla semplice espressione matematica della $f(x)$ in questo caso, possiamo facilmente scrivere le due equazioni di Eq. 38, ottenendo

$$\frac{\partial [\sum_i (y_i - a - bx_i)^2]}{\partial a} = \quad (39)$$

$$= -2 \sum_i (y_i - a - bx_i) = 0, \quad (40)$$

e

$$\frac{\partial [\sum_i (y_i - a - bx_i)^2]}{\partial b} = \quad (41)$$

$$= -2 \sum_i x_i (y_i - a - bx_i) = 0. \quad (42)$$

Lavorando di semplice algebra, si ha dalla prima

$$aN + b \sum_i x_i = \sum_i y_i \quad (43)$$

e dalla seconda

$$a \sum_i x_i + b \sum_i x_i^2 = \sum_i x_i y_i, \quad (44)$$

dove N è il numero di coppie di dati sperimentali, cioè il valore massimo che può assumere l'indice i delle sommatorie. Nonostante l'apparente complessità, quello appena scritto è un sistema lineare di due equazioni algebriche,

che porta alle soluzioni

$$a = \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{N \sum_i x_i^2 - (\sum_i x_i)^2} \quad (45)$$

$$b = \frac{N \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{N \sum_i x_i^2 - (\sum_i x_i)^2} \quad (46)$$

L'estensione al best-fit del minimo χ^2 è piuttosto immediata dal punto di vista concettuale, anche se un po' più complicata come scrittura. Stavolta nelle sommatorie delle Eqq. 39-42 compaiono dei termini $w_i = 1/(\Delta y_i)^2$ a moltiplicare, per cui le Eqq. 43 e 44 diventano

$$a \sum_i w_i + b \sum_i x_i w_i = \sum_i y_i w_i \quad (47)$$

$$a \sum_i x_i w_i + b \sum_i x_i^2 w_i = \sum_i x_i y_i w_i. \quad (48)$$

Anche senza ripetere il procedimento nei dettagli, si vede come le soluzioni abbiano la stessa "struttura" di Eqq. 45-46, solo che in tutti i termini delle sommatorie compare a moltiplicare un w_i e che N viene rimpiazzato da $\sum_i w_i$. Quindi, in sostanza, si ottengono le soluzioni espresse nelle Eqq. 22-27 (in quelle espressioni il denominatore delle frazioni è stato indicato con Δ'). Infine, per quanto riguarda la stima sulle incertezze dei parametri, Δa e Δb , esse possono essere determinate, al costo di diverse paginate di passaggi, applicando le regole di propagazione dell'errore alle soluzioni appena determinate.

[1] Occorre naturalmente distinguere tra nome delle grandezze e loro unità di misura. Per indicare l'unità di misura si usano diverse convenzioni, per esempio mettendone l'espressione simbolica tra parentesi quadre usando caratteri normali (non corsivi). Quando possibile, conviene usare prefissi moltiplicativi (M, k, m, μ , n, etc.). Nel caso in cui l'unità di misura sia arbitraria, si può usare [arb.un.] (il termine [a.u.] indica le unità atomiche). Naturalmente, se l'unità di misura non c'è (grandezze adimensionali che non hanno unità di misura), essa non va espressa, e le eventuali parentesi quadre non devono comparire. Lo stesso si deve fare quando la grandezza rappresentata è "normalizzata", cioè ottenuta dal rapporto tra la grandezza (per esempio misurata) e una grandezza di riferimento, che deve avere le stesse dimensioni e unità di misura della prima.

[2] Per motivi oscuri, molto spesso la scelta di default per il character size dei valori sugli assi è tale che i caratteri non possono essere apprezzati in una stampa.

[3] In alcuni casi è invece conveniente eseguire una *linearizzazione dei dati*, cioè una manipolazione matematica dei dati di partenza che permetta di rendere lineare il loro andamento. Questa operazione, che deve essere accompagnata da opportune manipolazioni delle barre di errore, è ben diversa rispetto a quella che prevede l'uso della rappresentazione logaritmica o semi-logaritmica.

[4] L'indirizzamento dei files è sempre un'operazione "delicata", che può dare luogo a numerosi problemi. Ricorda-

te sempre che il computer cerca di default i files nella directory all'interno della quale Python viene lanciato. Non sempre è semplice capire quale sia questa directory: usando un editore interattivo, per esempio Pyzo, la directory può essere conosciuta digitando `pwd` nella riga di comando. Per navigare tra le directories, può fare comodo usare il comando `cd` seguito ad esempio da `pip` se si vuole andare nella sub-directory `pip`, da `\pip`, se si vuole salire di livello e passare alla directory `pip`, oppure da `..` se si vuole salire di livello, e basta.

[5] A causa di bachi presenti nelle distribuzioni di Pyzo e dei sistemi operativi, è possibile che la produzione di un file in un formato grafico che risulti correttamente stampabile richieda delle accortezze particolari, delle quali verrete a conoscenza con la pratica.

[6] L'analisi delle sorgenti di errore e delle incertezze da associare alle misure di grandezze elettriche effettuate con strumenti standard sarà oggetto di approfondite considerazioni più avanti in questo corso.

[7] Un esempio clamoroso di correlazione completa è quello in cui il parametro τ di un decadimento esponenziale viene espresso come prodotto di altri due parametri che non compaiono in altre parti della funzione di best-fit. I due parametri in questione sono ovviamente del tutto (anti)correlati, e il best-fit realizzato lasciando liberi tali due parametri può dare risultati del tutto inaffidabili.

[8] I.G. Hughes and T.P.A. Hase, *Measurements and their uncertainties* (Oxford University Press, Oxford, 2013).

- [9] Che la covarianza negativa implichi una “diminuzione” nell’incertezza della previsione è diretta conseguenza del fatto che la funzione di best-fit è esprimibile come una somma. Espressioni diverse si trovano in altri casi, come si può facilmente determinare calcolandone la covarianza.
- [10] L’argomento di questa sezione è stato messo in luce grazie alle utili e brillanti osservazioni di due vostri colleghi, FLD e AC, dell’anno 2015/16. A loro va un enorme ringraziamento.
- [11] L’opzione è implementata solo a partire da una certa versione di Python, o delle sue librerie. Dunque se incappate in un errore di compilazione relativo a questa opzione, vuol dire semplicemente che essa non è disponibile nella versione di Python che state impiegando. In questo caso, l’opzione è fissa al valore **False**.
- [12] Spesso è possibile compiere degli sforzi finalizzati a migliorare questa situazione. Per esempio, come vedremo in futuro, nelle misure automatizzate con Arduino si può in genere prevedere la possibilità di estrarre la deviazione standard sperimentale su un campione ragionevolmente esteso di misure ripetute in automatico. Questo è il punto di partenza per compiere una distinzione tra contributi stocastici e sistematici all’errore. Si potrebbe ad esempio eseguire gli eventuali best-fit per l’analisi dei dati usando solo l’incertezza stocastica e poi tenere conto del contributo sistematico nella conversione dei risultati del best-fit a unità “fisiche”. Avremo modo di trattare questo argomento in seguito.
- [13] Ho fatto un rapido controllo per verificare quale sia la definizione di errore sui parametri di best-fit in uso in alcuni software di trattamento dati (nei limiti delle mie possibilità). Usano l’asymptotic error, corrispondente all’opzione **False** di Python; MATLAB (R2017b), gnuplot (5.0.0), Mathematica (6.0, in forma piuttosto complicata); usa lo standard error, corrispondente all’opzione **True** di Python, IgorPro (versioni 6 e seguenti), un software orientato più a trattare dati di interesse per la fisica che non per l’ingegneria. OriginPro, che non uso, dovrebbe consentire la scelta dei due metodi attraverso opportuna definizione dell’incertezza di misura sui dati da fissare e similmente dovrebbero comportarsi, per quanto ne so, altri software pensati per analisi di dati in fisica.
- [14] P.R. Bevington and D.K. Robinson, Data reduction and error analysis for the physical sciences (McGraw-Hill, New York, 2002).
- [15] P.H. Richter, “Estimating Errors in Least-Squares Fitting”, TDA Progress Report 42-122 (1995); http://ipnpr.jpl.nasa.gov/progress_report/42-122/122E.pdf
- [16] <https://www.astro.rug.nl/software/kapteyn/kmpfittutorial.html>

Resistenza interna del generatore e best-fit

francesco.fuso@unipi.it

(Dated: version 5 - FF, 17 ottobre 2019)

Questa breve nota riporta i principali risultati che io ho ottenuto nell'esercitazione pratica della stima della resistenza interna del generatore di d.d.p. (resistenza di Thévenin) tramite best-fit. Gli esempi qui illustrati hanno lo scopo primario di mostrare cosa si può ottenere e come si può procedere in questa esercitazione pratica. Inoltre essi permettono di commentare brevemente alcuni aspetti potenzialmente rilevanti per l'interpretazione dei dati.

I. ESERCITAZIONE PRATICA E MODELLI

L'esercitazione, semplicissima sia dal punto di vista concettuale che da quello pratico, prevede di montare il circuito di Fig. 1(a), costituito da un generatore di differenza di potenziale V_0 (misurata a “circuito aperto”), un resistore di resistenza R_j appartenente a un set preliminarmente misurato con tester digitale configurato come ohmetro, e un amperometro che misura la corrente I_j che fluisce nel circuito. In questo esempio si impiega a questo scopo il tester digitale configurato come amperometro.

Sulla base delle conoscenze che abbiamo, possiamo distinguere almeno tre “livelli” (crescenti) di accuratezza nel modello che descrive il circuito:

- il modello più semplice, che si riferisce proprio alla Fig. 1(a), prevede che l'*unico* elemento resistivo sia costituito da R_j . Pertanto la legge che descrive questo modello è

$$I = \frac{V_0}{R}, \quad (1)$$

che stabilisce una proporzionalità inversa tra l'intensità di corrente I misurata dall'amperometro e la resistenza R ;

- un modello più raffinato impone di considerare la *resistenza interna* r_G che descrive il generatore di d.d.p. (reale) nell'approccio di Thévenin; poiché questa resistenza è in serie al circuito [vedi Fig. 1(b)], la legge diventa

$$I = \frac{V_0}{R + r_G}; \quad (2)$$

- un ulteriore affinamento del modello prevede di includere *anche* la resistenza interna r_A dell'amperometro (reale), secondo quanto schematizzato in Fig. 1(c); in questo caso la legge recita

$$I = \frac{V_0}{R + r_G + r_A}. \quad (3)$$

Verificheremo a posteriori che effettivamente la resistenza interna del generatore r_G ha un ruolo più rilevante, almeno per l'esempio qui considerato, della resistenza interna dell'amperometro r_A . Il valore nominale di r_A può

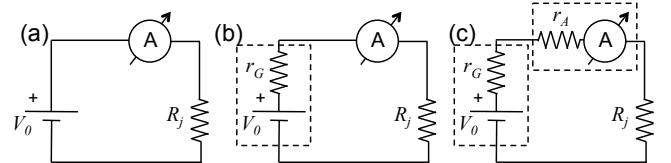


Figura 1. Rappresentazione schematica del circuito considerato (a), con esplicitate la resistenza interna r_G del generatore (b) e r_A dell'amperometro (c). I box tratteggiati racchiudono il generatore di d.d.p. (reale) e l'amperometro (reale).

essere dedotto dalle informazioni riportate nel manuale del tester (digitale), mentre a priori è del tutto sconosciuto il valore r_G . Di conseguenza possiamo interpretare questa esperienza come un modo, un po' involuto, per *misurare indirettamente* r_G , con il vantaggio di sfruttare tanti dati e il best-fit per aumentare l'accuratezza della nostra misura.

A. Resistenza interna dell'amperometro

Il manuale del tester digitale indica in $\Delta V_{\text{ins,fs}} = 200$ mV la caduta di potenziale per inserzione dello strumento usato come amperometro quando la lettura va a fondo scala. Questo dato è riportato senza incertezza (dato *nominal*e). La resistenza interna può essere dedotta dalla legge di Ohm come $r_A = \Delta V_{\text{ins,fs}} / I_{\text{fs}}$, dove I_{fs} è il *fondo scala* della portata di misura in corrente che si sta effettuando. L'esperienza richiede di utilizzare resistenze R_j in un vasto intervallo di valori (oltre sei decadi), in corrispondenza dei quali anche l'intensità di corrente che fluisce nel circuito varia di altrettanti ordini di grandezza. Allo scopo di mantenere la significatività delle misure, cioè di avere un numero adeguato di cifre significative, nell'esperienza è necessario usare parecchie portate per la misura di corrente, a cui corrispondono diversi valori di I_{fs} e quindi di r_A .

In particolare si ottiene nominalmente $r_A = \{1, 10, 10^2, 10^3, 10^4\}$ ohm per i fondi scala $\{200, 20, 2, 0.2 \text{ mA}\}$ e $20 \mu\text{A}$, rispettivamente. Fortunatamente, grazie alla dipendenza inversa della corrente con la resistenza espresso dalle leggi scritte prima, i valori più alti di resistenza interna corrispondono ai fondo scala che si usano quando la resistenza R_j è grande. Questo

ci autorizza a trascurare, solo in prima battuta e salvo ulteriori considerazioni, l'effetto della resistenza interna dell'amperometro nel circuito sotto esame.

II. DATI E BEST-FIT

Poiché l'esercitazione pratica richiede di variare R usando j distinti resistori con resistenza R_j , appartenenti a un vasto intervallo, e poiché anche l'intensità di corrente corrispondente, I_j , varia nello stesso intervallo, è opportuno rappresentare i dati in scala *logaritmica*. In Python questo si ottiene per esempio con i comandi `pylab.xscale('log');` `pylab.yscale('log')`. Osservate che, qualora valesse la legge espressa da Eq. 1, i dati rappresentati in scala logaritmica dovrebbero essere tutti allineati lungo una retta con coefficiente angolare -1 (sulla scala del grafico). Per agevolare l'individuazione a occhio di questa direzione, è opportuno che i due assi del grafico siano aggiustati in modo da coprire lo stesso intervallo in decadi (per esempio sette decadi). Dunque la retta cercata ha la direzione della “*bisettrice*” del grafico così prodotto (la “*bisettrice*” che va da sinistra in alto a destra in basso).

Se si osservano i dati riportati in Fig. 2 risulta evidente che essi *non* seguono (tutti) l'andamento previsto da Eq. 1: in particolare, per bassi valori di R_j la corrente è evidentemente più bassa di quanto ci si aspetterebbe secondo quella legge. La presenza delle resistenze interne può correttamente interpretare l'osservazione sperimentale. La linea continua sovrapposta ai dati di figura è il risultato di un best-fit *numerico* secondo la funzione

$$I = \frac{V_0}{R + r}, \quad (4)$$

con $r = (r_G + r_A)$, dove i parametri di best-fit sono V_0 e $[1]$. Notiamo che, ovviamente, *non è possibile* pretendere che l'algoritmo di best-fit distingua tra le due resistenze interne (in altre parole, r_A e r_G sono completamente anti-correlati tra loro, e il best-fit deve considerare un “unica” resistenza interna somma delle due). Dal punto di vista concettuale questa scelta presenta un problema: infatti sappiamo che r_A *non* è costante per tutte le misure, per cui a rigore non ha senso considerarlo un parametro per l'intero set di dati. Torneremo in seguito su questo aspetto.

Il best-fit è stato eseguito modificando (in modo molto molto leggero) lo script già impiegato e discusso per l'Esercitazione “zero”. A parte dettagli (rappresentazione logaritmica, nomi degli assi, scale), la modifica di maggior rilievo è la definizione della funzione, che in questo caso è data da queste due linee di script:

```
def ff(x, aa, bb):
    return aa/(bb+x)
```

dove `aa` e `bb` hanno il ruolo dei parametri “fisici” (correttamente dimensionati) V_0 e r ; i valori iniziali di tali

parametri, che devono essere forniti alla routine di minimizzazione, possono facilmente essere dedotti dall'analisi del circuito: V_0 è infatti atteso dell'ordine della d.d.p. misurata a circuito aperto e r è attesa dell'ordine della decina di ohm. Per il best-fit di figura sono state considerate solamente le incertezze ΔI_j sulle correnti (in seguito si tornerà anche su questo punto) e, vista l'origine prevalentemente sistematica (errore di calibrazione) delle incertezze, è stata usata l'opzione `absolute_sigma = False`. I risultati del best-fit sono

$$V_0 = (4.94 \pm 0.02) \text{ V} \quad (5)$$

$$r = (18.6 \pm 0.6) \text{ ohm} \quad (6)$$

$$\chi^2/\text{ndof} = 54/13 \quad (7)$$

$$\text{norm.cov.} = 0.38 \quad (8)$$

$$\text{absolute_sigma} = \text{False}. \quad (9)$$

Il valore di V_0 non è compatibile con la misura a circuito aperto eseguita collegando il solo tester digitale, configurato come voltmetro, al generatore di d.d.p.: $V_{0,ap} = (5.04 \pm 0.03)$ V. La correlazione positiva che abbiamo trovato è attesa per come è scritta la funzione (l'aumento di V_0 è “compensato” da un aumento di r) e il χ^2 ridotto è nettamente superiore all'unità.

Prima di proseguire con altre varianti di best-fit, notiamo, con un ragionamento a spanne, come i dati suggeriscano che la resistenza interna dell'amperometro non possa essere l'(unica) responsabile per gli effetti registrati. Infatti i punti sperimentali che maggiormente si discostano dall'andamento rettilineo (in carta logaritmica) sono stati acquisiti usando il fondo scala 200 mA per l'amperometro, a cui corrisponde una resistenza interna $r_A = 1$ ohm (nominale). Essendo il valore di r ottenuto dal best-fit superiore di un ordine di grandezza, si può ragionevolmente supporre che la resistenza interna del generatore giochi il ruolo predominante.

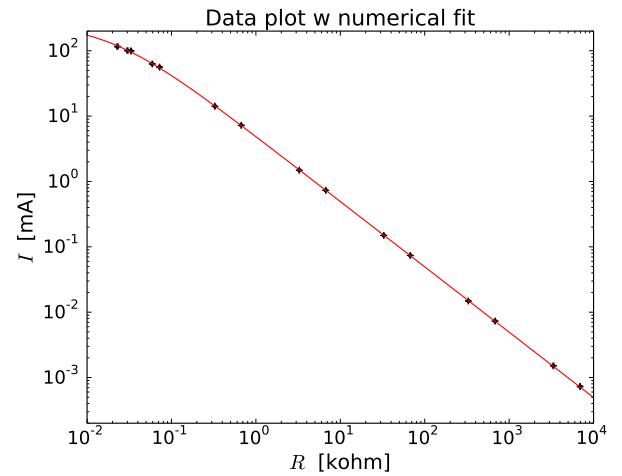


Figura 2. Dati e best-fit numerico secondo la funzione di Eq. 4. Notate che le barre di errore sono state correttamente incluse nel grafico, ma esse sono tali da essere difficilmente visibili.

A. Incertezze ΔR_j

La misura delle resistenze R_j è eseguita con il tester (digitale), per cui essa è inevitabilmente affetta da un errore normalmente dominato dall'incertezza di calibrazione. Almeno in linea di principio, le incertezze ΔR_j non possono essere considerate trascurabili a priori. Di esse possiamo tenere conto nel modo (non elegante e non sicuro) che abbiamo già discusso, basato sulla propagazione dell'errore.

L'“errore efficace”, o “errore equivalente”, $\Delta I_j|_{\Delta R_j} = \Delta R_j(\partial I/\partial R)$ può essere determinato facilmente derivando l'Eq. 4:

$$\Delta I_j|_{\Delta R_j} = \Delta R_j \frac{V_0}{(R_j + r)^2}. \quad (10)$$

Questo errore va quindi sommato in quadratura con ΔI_j , allo scopo di ottenere il peso che compare nella somma dei residui, ovvero nella determinazione della grandezza che deve essere minimizzata nel fit. È evidente che per impiegare questo metodo è necessaria la conoscenza di V_0 e r i quali, essendo parametri nel best-fit, devono essere preliminarmente stimati con una procedura in cui l'incertezza ΔR_j è trascurata: per noi, essi sono quindi quelli scritti nelle Eqq. 5, 6.

Il nuovo best-fit così prodotto è riportato in Fig. 3 assieme al grafico dei residui normalizzati. Si osserva come, nonostante vengano qui considerati anche gli errori ΔR_j , i residui normalizzati siano spesso al di fuori del range atteso (compreso tra -1 e +1), in particolare per bassi valori di R_j . I risultati del best-fit sono:

$$V_0 = (4.95 \pm 0.02) \text{ V} \quad (11)$$

$$r = (18.6 \pm 0.6) \text{ ohm} \quad (12)$$

$$\chi^2/\text{ndof} = 26/13 \quad (13)$$

$$\text{norm.cov.} = 0.49 \quad (14)$$

$$\text{absolute_sigma} = \text{False}. \quad (15)$$

Essi sono in accordo con quanto trovato in precedenza, a parte un'ovvia diminuzione del χ^2 (stavolta $\chi^2_{\text{red}} \simeq 2$) e un diverso valore di covarianza normalizzata, la cui origine può essere ascritta al diverso peso attribuito nella procedura di best-fit ai vari dati dovuto alla diversa determinazione dell'incertezza.

B. Inclusione di r_A

Un motivo per cui la “qualità” dei best-fit ottenuti non è particolarmente elevata, in particolare per quanto riguarda il valore di V_0 diverso da quello misurato a circuito aperto, può essere individuato nel trattamento che abbiamo riservato a r_A . Come già affermato, tale parametro non è costante per l'intero set di dati, ma dipende dalla portata effettivamente impiegata per la misura delle correnti in corrispondenza dei vari valori R_j . Visto che conosciamo a priori, almeno nominalmente, il valore di

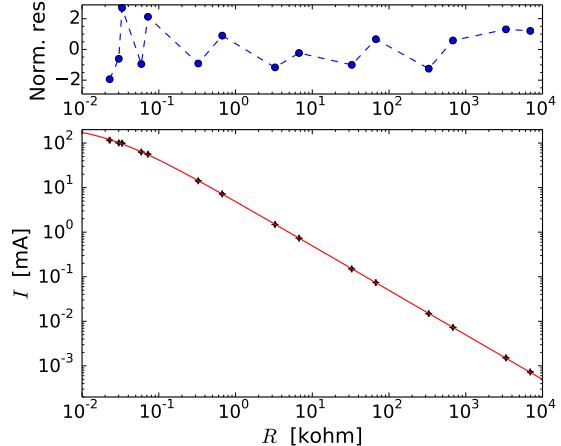


Figura 3. Analogico di Fig. 2 con il best-fit eseguito tenendo conto anche dell'incertezza ΔR_j (attraverso propagazione dell'errore); il pannello superiore mostra il grafico dei residui normalizzati.

r_A per ciascuna delle portate impiegate, una via semplice per tenerne conto può essere quella di rimpiazzare R_j con $X_j = R_j + r_A$. In questo modo la funzione di best-fit diventa

$$Y = \frac{V_0}{X + r_G}, \quad (16)$$

che contiene solo la resistenza interna r_G , supposta costante per l'intero set di dati.

L'ulteriore manipolazione a cui vengono sottoposti i dati può essere facilmente realizzata con Python, per esempio creando un'ulteriore colonna del nostro file di testo che contiene r_A nelle effettive condizioni di misura (portata). Facciamo subito una considerazione importante: questa manipolazione *dovrebbe* essere accompagnata da un aumento dell'incertezza, cioè dovremmo porre, ad esempio, $\Delta X_j = \sqrt{(\Delta R_j)^2 + (\Delta r_A)^2}$. Poiché, però, Δr_A non è nota, usiamo i valori nominali accettando di conseguenza una possibile sottostima dell'incertezza.

L'esito dell'operazione è riportato in Fig. 4: si nota una piccola riduzione del valore medio dei residui normalizzati rispetto a Fig. 3, che si riflette in un'ulteriore riduzione del χ^2 . I risultati sono infatti

$$V_0 = (5.05 \pm 0.02) \text{ V} \quad (17)$$

$$r_G = (18.6 \pm 0.4) \text{ ohm} \quad (18)$$

$$\chi^2/\text{ndof} = 16/13 \quad (19)$$

$$\text{norm.cov.} = 0.49 \quad (20)$$

$$\text{absolute_sigma} = \text{False}; \quad (21)$$

il più interessante è il valore di V_0 che questa volta è compatibile con $V_{0,ap} = (5.04 \pm 0.03)$ V. Si noti che, d'altra parte, il valore di r_G qui ottenuto è compatibile con quello di $r = r_G + r_A$ di Eq. 12, a dimostrazione che qui la resistenza interna dell'amperometro è di fatto trascura-

bile in termini numerici ai fini della determinazione del corrispondente parametro di best-fit.

Il grafico dei residui normalizzati non mostra grandi differenze qualitative rispetto a quello determinato in precedenza: osservando che le maggiori discrepanze rispetto allo zero si hanno per bassi valori di R_j , ovvero alte correnti, si può ipotizzare una spiegazione fisica i cui effetti non sono contenuti nel modello impiegato. Per intensità di corrente di parecchie decine di mA, la potenza dissipata dagli elementi resistivi, in particolare da R_j (e anche dal fusibile di cui è dotato il generatore, si veda dopo) può diventare non trascurabile. Per esempio, il valore massimo della potenza Joule $P_J = RI^2$ risulta, per i dati qui considerati, superiore a 0.3 W (che è anche maggiore della massima potenza sopportabile da alcuni dei resistori in uso in laboratorio, il cui valore nominale è 1/4 W). Per tali potenze è probabile che si verifichi un surriscaldamento dei componenti accompagnato da una variazione di R_j , effetto che è ovviamente trascurato nel nostro modello [2].

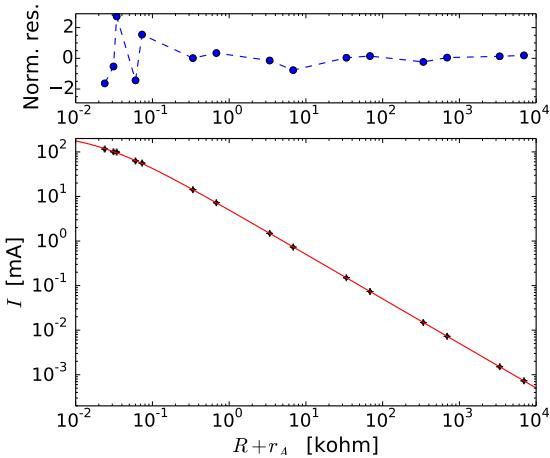


Figura 4. Analogo di Fig. 3, ma considerando per l'asse orizzontale la grandezza $X = R + r_A$, ed eseguendo un best-fit di conseguenza, secondo quanto discusso nel testo.

C. “Banda di confidenza”

Questa sezione descrive un’ulteriore operazione sul best-fit la cui utilità è spesso dubbia (si tratta di una sorta di “abbellimento”, che in genere va considerato come opzionale). L’operazione è finalizzata a disegnare sul grafico altre due curve che rappresentano il limite superiore e inferiore delle previsioni del modello. Chiamiamo l’intervallo compreso tra queste due curve “banda di confidenza”, denominazione che è a rigore corretta solo nel caso in cui il best-fit sia condotto utilizzando un “vero” χ^2 , cioè, in sostanza, quando i dati di partenza hanno un’incertezza di carattere prevalentemente stocastico. In queste condizioni, come sapete dallo scorso anno, i da-

ti contenuti all’interno della banda rappresentano una previsione valida con confidenza del 95% [3].

Disegnare le due curve significa valutare la “massima sovrastima” e la “massima sottostima” della previsione sulla base della funzione modello prescelta, che quindi deve essere ritenuta valida per la descrizione dei risultati sperimentali nell’intervallo di dati considerato. La sovrastima o sottostima della previsione è dovuta alla circostanza che i parametri della funzione modello sono determinati dal best-fit con un certo errore. Inoltre, nel caso in cui ci sia più di un parametro, sappiamo che la covarianza tra i parametri deve anche essere tenuta in conto.

Ricordiamo infatti che, in generale, per una funzione $f(\alpha, \beta)$ di due variabili aleatorie α, β , si ha, al primo ordine e con ovvio significato dei simboli:

$$\sigma_f^2 = \sigma_\alpha^2 \left(\frac{\partial f}{\partial \alpha} \right)^2 + \sigma_\beta^2 \left(\frac{\partial f}{\partial \beta} \right)^2 + 2\sigma_{\alpha\beta} \frac{\partial f}{\partial \alpha} \frac{\partial f}{\partial \beta}. \quad (22)$$

Facendo riferimento al nostro esperimento, in particolare al best-fit eseguito secondo la funzione di Eq. 16, α e β sono i due parametri V_0 e r_G (supponiamo qui che i parametri del best fit possano essere considerati come variabili aleatorie); le varianze e covarianze $\sigma_\alpha^2, \sigma_\beta^2, \sigma_{\alpha\beta}$ sono rispettivamente gli elementi C_{11}, C_{22} e C_{12} (ovvero C_{21} , la matrice è simmetrica) della *matrice di covarianza* generata dalla routine di minimizzazione. Infine, le derivate parziali che compaiono in Eq. 22 possono essere calcolate usando la funzione di Eq. 16.

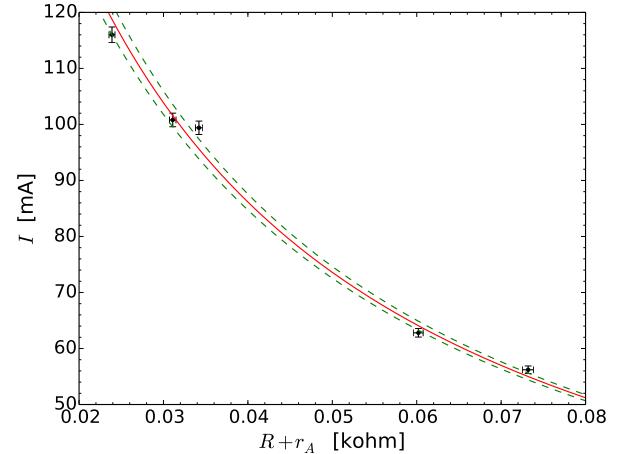


Figura 5. Rappresentazione in scala lineare di un subset dei dati con sovrapposta la curva di best-fit (linea rossa continua, analoga a quella mostrata in Fig. 4) e le “curve di confidenza” (linee verdi tratteggiate) costruite come discusso nel testo.

In buona sostanza, e salvo errori (sempre possibili in questi casi), si ottiene, per il caso che stiamo considerando e usando la simbologia introdotta,

$$\sigma_f^2 = C_{11} \left(\frac{1}{X + r_G} \right)^2 + C_{22} \left(\frac{V_0}{(X + r_G)^2} \right)^2 - \quad (23)$$

$$- 2C_{12} \frac{1}{X + r_G} \frac{V_0}{(X + r_G)^2}. \quad (24)$$

La radice quadrata di questa varianza, calcolata per i valori X usando i parametri V_0 e r_G determinati dal best-fit, rappresenta l'*incertezza sulla previsione*. Le curve che stiamo cercando possono quindi essere costruite aggiungendo e sottraendo dalla curva di best-fit il valore $\sqrt{\sigma_f^2}/2$ determinato per ogni valore della “variabile indipendente” X .

Per i dati considerati in questa nota, l’incertezza sulla previsione risulta piccola rispetto al valore della previsione stessa: graficando i dati e le funzioni nello stile impiegato finora, le “curve di confidenza” risultano del tutto indistinguibili da quella di best-fit (questo è un buon motivo pratico per cui in genere non vale la pena di calcolarsi la banda di confidenza). Per il solo scopo di illustrare in maniera tipograficamente chiara quello che si ottiene, la Fig. 5 mostra un subset dei dati che comprende solo i valori più bassi di resistenza esplorati nell’esperimento, dove le discrepanze sono più rilevanti per le motivazioni a cui accenneremo in seguito, in rappresentazione lineare. Si vede come almeno un dato sperimentale cada al di fuori della “banda di confidenza”, presumibilmente per il verificarsi dei fenomeni di surriscaldamento a cui abbiamo già fatto cenno.

III. LINEARIZZAZIONE DELL’ANDAMENTO

In questo paragrafo utilizziamo una tecnica di linearizzazione dei dati in modo da ricondurci a un best-fit di tipo lineare ed eseguirlo con le formule analitiche a voi forse note dall’anno scorso, o che potete trovare nella nota di illustrazione alla “Esercitazione 0”. Una buona possibilità di linearizzazione è rappresentata dall’utilizzare, invece che la corrente I , il suo reciproco $Y = 1/I$. Infatti usando questa nuova grandezza (che avrà le dimensioni del reciproco di una intensità di corrente) l’Eq. 4 diventa

$$Y = \frac{R}{V_0} + \frac{r}{V_0} \quad (25)$$

che è proprio l’equazione di una retta $y = a + bx$, con $V_0 = 1/b$ e $r = aV_0 = a/b$.

Questa linearizzazione implica necessariamente di manipolare i dati, operazione che può essere fatta molto agevolmente da Python. Inoltre occorre anche manipolare le incertezze ΔI_j in modo da esprimere, con le ben note regole di propagazione dell’errore, l’incertezza ΔY_j . Una volta eseguite queste operazioni, il best-fit può essere condotto in modo analitico. L’esito è mostrato in Fig. 6, dove, come ulteriori affinamenti della procedura, sono state considerati gli errori ΔR_j (introducendo il debito “errore efficace”) e la resistenza interna dell’amperometro secondo quanto stabilito in precedenza. I risultati del best-fit, riportati ai parametri di interesse fisico (e facendo opportuno uso della propagazione dell’errore

massimo), sono:

$$V_0 = (5.05 \pm 0.01) \text{ V} \quad (26)$$

$$r_G = (18.6 \pm 0.4) \text{ ohm} \quad (27)$$

$$\chi^2/\text{ndof} = 16/13, \quad (28)$$

dove, coerentemente con l’uso della procedura analitica, non esprimiamo né la covarianza normalizzata, né l’opzione per la determinazione dell’incertezza. Osservate che, in questo caso in cui $\chi^2_{rid} \simeq 1$, l’errore sui parametri fornito dalla procedura analitica, che corrisponde all’uso dell’opzione `absolute_sigma = True` nella routine di minimizzazione numerica, è lo stesso di quello che si ottiene usando `absolute_sigma = False`, come fatto in precedenza.

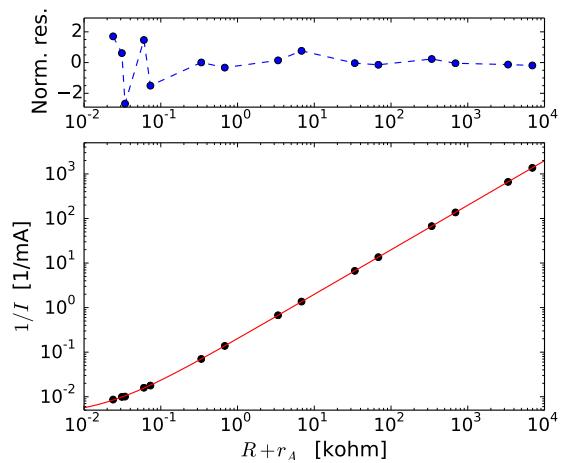


Figura 6. Dati linearizzati secondo quanto discusso nel testo e best-fit analitico secondo la funzione di Eq. 25. Notate che le barre di errore sono state anche in questo caso correttamente incluse nel grafico, ma esse sono tali da essere difficilmente visibili. Nel best-fit sono stati considerati gli errori ΔR_j e la resistenza interna nominale dell’amperometro r_A ; il pannello superiore riporta il grafico dei residui normalizzati.

IV. MISURA “ALLA THÉVENIN” E COMMENTI

Secondo il cosiddetto *teorema, o modello, di Thévenin*, qualsiasi generatore di d.d.p. (che comprenda solo elementi passivi) può essere modellato come un generatore ideale di d.d.p. V_{Th} con in serie una resistenza R_{Th} . Il modello prevede delle regole pratiche per determinare V_{Th} e R_{Th} sia nel caso in cui sia noto il circuito sotto esame (questa sarà la situazione che incontrerete in prossime esercitazioni) che in quello in cui il generatore reale è, di fatto, una *scatola nera*. Questa è la situazione della presente esperienza.

In tali condizioni V_{Th} può essere *misurato* collegando al generatore un voltmetro ideale. Tenendo conto che, grazie alla sua grande resistenza interna ($r_V = 10 \text{ Mohm}$,

nominali), il tester digitale approssima piuttosto bene un voltmetro ideale, si ha sostanzialmente $V_{Th} = V_{0,ap}$ (l'aggettivo “aperto” qui richiamato nel pedice fa proprio riferimento al fatto che il tester digitale configurato come voltmetro si comporta come un *circuito aperto*, cioè non passa praticamente corrente attraverso di esso). La resistenza di Thévenin R_{Th} è invece identificabile con r_G . Secondo la ricetta di Thévenin, essa può essere *misurata* collegando un carico resistivo R_L esterno, misurando la d.d.p. ai capi del carico con un voltmetro ideale, e, in definitiva, applicando la legge di Ohm al partitore di tensione costituito dalla serie di r_G e R_L . Nella ricetta si specifica anche l'impiego di una R_L simile, in valore, a r_G . Questo metodo consente di trascurare la resistenza interna dell'amperometro (non c'è amperometro nel circuito) e dunque di disinteressarsi della sua valutazione.

Nell'esempio qui considerato è stato impiegato il valore $R_L = (22.9 \pm 0.4)$ ohm (una tale resistenza si è realizzata mettendo in parallelo i resistori di valore nominale 33 e 68 ohm). In queste condizioni la d.d.p. misurata ai capi di R_L è $\Delta V_L = (2.74 \pm 0.02)$ V. Poiché nel partitore è $V_L = V_{0,ap}R_L/(R_L + R_{Th})$, da cui $R_{Th} = R_L(V_{Th}/V_L - 1)$, si ottiene, usando in modo opportuno la propagazione dell'errore (in questo caso non conviene usare la propagazione dell'errore relativo), $R_{Th} = (19.2 \pm 0.9)$ ohm. Questo valore è in accordo entro l'incertezza con quello ottenuto per r_G dai best-fit, ma, ovviamente, è affetto da un errore più rilevante, essendo

determinato da una singola misura e non dall'intero set dei dati.

Come ultima osservazione torniamo sugli aspetti “fisici” che sono presumibilmente coinvolti nell’esperienza, partendo proprio dalla clausola (presenza di soli componenti passivi nel generatore) che è inserita nel modello di Thévenin. In effetti l'alimentatore usato come generatore di d.d.p. non contiene al suo interno solo elementi passivi. Infatti questo alimentatore è di tipo *switching* (avremo probabilmente modo di accennare al suo funzionamento, molto “smart”, in futuro) ed è realizzato con parecchia circuiteria elettronica comprendente certamente degli elementi “attivi” (transistors). Dispositivi di questo tipo sono caratterizzati da resistenze interne molto basse e potenzialmente dipendenti dalle condizioni di operazione, in particolare dalla richiesta di corrente. Dunque è possibile che la nostra descrizione sia inadeguata.

Inoltre l'alimentatore dispone di un fusibile (corrente massima nominale di 100 mA) posto in serie alla boccola di uscita. Questo componente è sicuramente resistivo (il fusibile è una resistenza che eventualmente fonde surriscaldandosi per effetto Joule) e in effetti la sua resistenza è dell'ordine della r_G , ovvero R_{Th} , ottenuta nell'esperienza (kudos to Diego per l'illuminante precisazione). Dunque la valutazione della resistenza interna del generatore eseguita nell'esperienza è, in gran parte, influenzata dalla presenza di questo elemento resistivo che noi abbiamo considerato parte integrante del generatore di d.d.p..

- [1] Affinché la curva del fit sia graficata in modo visivamente accettabile, è bene che il valore della funzione di fit venga calcolato su un array di punti *equispaziati logaritmicamente*. In Python questo si può ottenere con il comando `xx=numpy.logspace(-2,4,1000)`, che appunto crea un array (denominato qui `xx`) di 1000 punti equispaziati logaritmicamente tra 10^{-2} e 10^4 (state attenti alla sintassi del comando).
- [2] Per limitare gli effetti del surriscaldamento conviene eseguire le misure, specialmente quelle che comportano un elevato passaggio di corrente, in tempi brevi.
- [3] Il fatto che le incertezze sperimentali sui dati di questo esempio non siano, presumibilmente, di carattere puramente statistico impedisce di definire quantitativamente la confidenza della previsione.

Misure rms con multmetro

francesco.fuso@unipi.it; <http://www.df.unipi.it/~fuso/dida>

(Dated: version 3 - FF, 11 novembre 2015)

Queste poche righe intendono richiamare alcuni concetti relativi alle grandezze periodiche alternate e fornire un'interpretazione (thanks, Diego) del comportamento che un multmetro ha nella lettura di valori rms per grandezze alternate. Questa seconda parte riguarda un argomento abbastanza "raffinato", probabilmente di interesse non generale, ma comunque utile per capire il perché di quanto osservato.

I. POTENZA MEDIA E VALORI RMS

L'introduzione dei valori di ampiezza rms (valori *efficaci*) per segnali alternati ha una chiara motivazione pratica. Supponiamo infatti di avere un elemento circuitale in cui scorre una corrente di intensità $I(t) = I_0 \cos(\omega t + \phi_1)$ e ai cui capi si misura una differenza di potenziale $V(t) = V_0 \cos(\omega t + \phi_2)$. In altre parole, sia d.d.p. che corrente hanno un *andamento sinusoidale*, evidentemente con media temporale nulla (grandezze *alternate*). Senza perdere di generalità, si può sempre immaginare di traslare l'origine dei tempi in modo tale che $\phi_1 = 0$, per cui si avrà $I(t) = I_0 \cos(\omega t)$ e $V(t) = V_0 \cos(\omega t + \Delta\phi)$, con $\Delta\phi = \phi_2 - \phi_1$ (*sfasamento*).

La potenza "instantanea" che "interessa" (è "dissipata" nel) l'elemento circuitale, ovvero quella, segno a parte, erogata dal generatore a cui l'elemento è collegato, si scrive

$$\begin{aligned} P(t) &= I(t)V(t) = I_0 V_0 \cos(\omega t) \cos(\omega t + \Delta\phi) = \\ &= I_0 V_0 \cos(\omega t)(\cos(\omega t) \cos(\Delta\phi) - \sin(\omega t) \sin(\Delta\phi)) \end{aligned} \quad (1)$$

dove abbiamo usato una nota relazione trigonometrica. Questa potenza oscilla periodicamente tra zero e il valore massimo $I_0 V_0$.

Dal punto di vista pratico, allo scopo di quantificare l'effettiva dissipazione di potenza da parte dell'elemento circuitale (o "utilizzatore") occorre determinare il valore *medio nel tempo* (d'ora in avanti solo *medio*) di $P(t)$. Trattandosi di una funzione periodica, la media si può fare integrando su un periodo T , cioè:

$$\langle P \rangle = \frac{1}{T} \int_{-T/2}^{T/2} P(t) dt. \quad (3)$$

Nell'integrazione della funzione di Eq. 1 i termini dispari, quelli del tipo $\cos(\omega t) \sin(\omega t)$, si annullano, poiché si tratta di funzioni a media nulla; rimangono da integrare solo i termini del tipo $\cos^2(\omega t)$. Si ottiene facilmente (al secondo anno di Università dovete saper fare questi integrali!) che, *per funzioni sinusoidali*:

$$\langle P \rangle = \frac{I_0 V_0}{2} \cos(\Delta\phi) \quad (4)$$

L'origine fisica dell'eventuale termine *di sfasamento* $\Delta\phi$ vi sarà chiara andando più avanti nel corso. Nel caso,

molto comune, che il componente considerato sia resistivo, cioè ohmico (una lampadina, per esempio), non c'è sfasamento tra corrente e tensione (si dice che il *fattore di potenza* $\cos(\Delta\phi)$ vale uno), per cui $\langle P \rangle = I_0 V_0 / 2 = V_0^2 / (2R) = RI_0^2 / 2$, dove per le ultime due uguaglianze si è usata la legge di Ohm.

Il valore rms di una grandezza periodica alternata $f(t)$, di periodo T , è definito come:

$$f_{rms} \equiv \sqrt{\langle f^2 \rangle} = \sqrt{\frac{1}{T} \int_{-T/2}^{T/2} f^2(t) dt}. \quad (5)$$

A. Forme d'onda sinusoidali

Sulla base di quanto stabilito, una forma d'onda sinusoidale per un "segnaletico" di tensione si può scrivere come $V(t) = V_0 \cos(\omega t)$, con V_0 ampiezza e ω frequenza angolare (o pulsazione) del segnale. Nell'espressione ho omesso il termine di "fase costante", supponendo di poter traslare l'origine dei tempi a piacere in modo da ottenere una dipendenza dal tempo di tipo coseno. Questo non fa perdere generalità, come potete facilmente verificare, ma semplifica di parecchio la matematica.

Supponiamo allora di avere una *forma d'onda sinusoidale*, cioè una funzione $f(t) = f_0 \cos(\omega t)$, è immediato verificare che si ha

$$f_{rms} = \frac{f_0}{\sqrt{2}} \text{ onda sinusoidale}, \quad (6)$$

Piccola osservazione pratica: f_0 rappresenta l'ampiezza e *non l'ampiezza picco-picco* f_{pp} della forma d'onda considerata! L'ampiezza picco-picco, cioè la distanza fra minimi e massimi di ampiezza, vale infatti $f_{pp} = 2f_0$. Nel caso della corrente di rete (quella fornita dall'ENEL), si sa che $V_{rms} = 230 - 240$ V. Provate a calcolare la corrispondente V_{pp} : troverete valori molto alti (oltre 500 V), che dovrebbero indurvi a porre particolare attenzione quando maneggiate apparecchi elettrici!

Sulla base di quanto dimostrato in precedenza, nel caso sinusoidale e con carichi resistivi si ha immediatamente

$$\langle P \rangle = V_{rms} I_{rms} = \frac{V_{rms}^2}{R} = RI_{rms}^2, \quad (7)$$

cioè otteniamo le stesse relazioni formali che valgono in corrente continua, a patto di considerare al posto delle

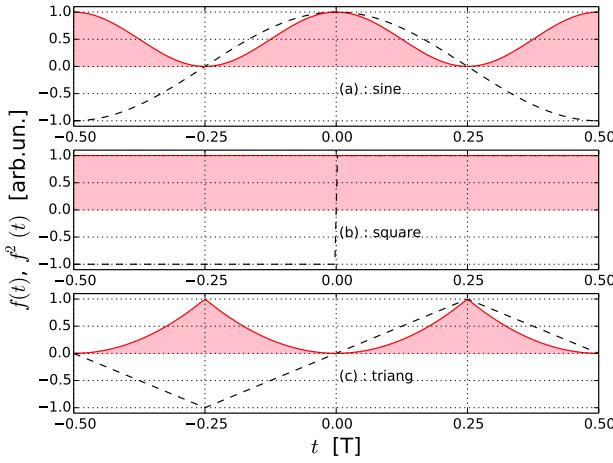


Figura 1. Grafici delle funzioni considerate nel testo per il caso di onda sinusoidale (a), onda quadra (b), onda triangolare (c). L'asse orizzontale copre un singolo periodo di oscillazione. Le curve tratteggiate nere sono le funzioni che rappresentano le forme d'onda, quelle rosse continue sono i quadrati (moduli quadrati) e la superficie riempita in rosa rappresenta l'area sottesa a queste ultime curve. La figura è ovviamente fatta con Python.

ampiezze V_0 e I_0 i valori rms V_{rms} , I_{rms} . L'affermazione di Eq. 7 è *del tutto generale*, cioè può essere applicata anche a forme d'onda diverse dalla sinusoidale, purché periodiche e alternate. Questo è il motivo per cui i valori rms (*root mean square*) hanno rilevanza pratica.

Può essere utile dare una rapida occhiata agli andamenti temporali delle funzioni che abbiamo considerato. La Fig. 1(a) mostra un'ipotetica forma d'onda sinusoidale $f(t) = f_0 \cos(\omega t)$ con ampiezza $f_0 = 1$ [arb.un.] (dunque ampiezza picco-picco $f_{pp} = 2$ [arb.un.]). L'asse delle ascisse corrisponde a un singolo periodo $T = 2\pi/\omega$. La curva nera tratteggiata rappresenta la funzione $f(t)$, la rossa continua è $f^2(t)$. Nella definizione di valore rms (Eq. 5) occorre calcolare l'integrale di quest'ultima funzione, che è proporzionale all'area sottesa alla curva (riempita di rosa in figura). Si vede come questa area, da prendere segnata, sia diversa da zero (e sempre positiva), mentre invece l'area sottesa a $f(t)$ è nulla, in accordo con il fatto che il segnale è alternato.

B. Forme d'onda quadre

Generalmente è sempre vero, almeno nei casi qui considerati, che $f_{rms} \propto f_0$. Per forme d'onda sinusoidali, come abbiamo appena analizzato, il coefficiente di proporzionalità è $1/\sqrt{2}$. L'obiettivo che ci prefiggiamo è ora quello di stabilire il coefficiente di proporzionalità per un'altra forma d'onda, per esempio per l'*onda quadra* (alternata, periodica e "simmetrica") rappresentata con la curva nera tratteggiata in Fig. 1(b).

Facendo il quadrato si ottiene una funzione costante di valore f_0^2 e l'area sottesa a questa curva è quella di un rettangolo. Ci vuole un attimo a verificare che, per una *forma d'onda quadra*:

$$f_{rms} = f_0 \text{ onda quadra .} \quad (8)$$

Dunque per un'onda quadra il fattore di proporzionalità è uno e il valore rms coincide con l'ampiezza.

C. Forme d'onda triangolari

Un altro caso interessante è quello dell'onda triangolare [Fig. 1(c)]. Anche stavolta il valore rms è non nullo e si vede che l'area sottesa alla curva $f(t)^2$ nel periodo è pari al quadruplo dell'area sottesa alla stessa funzione nel semiperiodo $t = (0, T/4]$.

Come si può facilmente verificare, in questo quarto di periodo si ha $f(t) = f_0 t / (T/4)$. Determiniamo allora f_{rms} per una *forma d'onda triangolare* (simmetrica):

$$f_{rms} = \sqrt{4 \times \frac{1}{T} \int_0^{T/4} \frac{f_0^2}{(T/4)^2} t^2 dt} = \quad (9)$$

$$= f_0 \sqrt{\frac{64}{T^3} \frac{T^3}{3 \times 64}} = \frac{f_0}{\sqrt{3}} \text{ onda triangolare .} \quad (10)$$

Dunque per un'onda triangolare il fattore di proporzionalità è $1/\sqrt{3}$.

II. MISURA RMS

Misurare il valore rms di una grandezza alternata è tutt'altro che banale. Nel futuro vedrete, spero vivamente, apparecchi che possono svolgere egregiamente questa funzione (ad esempio i "lock-in", o amplificatori a detezione sincrona). Sicuramente un multimetro portatile non contiene al suo interno raffinatezze che possano rendere sempre affidabile la valutazione, anche se il costruttore dichiara che è proprio il valore rms a essere restituito dallo strumento quando questo è impiegato per grandezze (tensioni e correnti) alternate.

Studiando un po' gli schemi (per esempio quello del multimetro analogico, che è disponibile, ma anche il digitale dovrebbe usare schemi simili), si vede come la misura rms avvenga in un modo diverso rispetto a quanto stabilito dalle definizioni matematiche, dato che non è semplice, in elettronica, costruire un segnale che sia il quadrato di un altro segnale.

A grandi linee, infatti, nei multimetri usati come voltmetri in alternata il segnale viene fatto passare attraverso un raddrizzatore (un diodo, ne parleremo più avanti) che in sostanza "taglia" (pone uguale a zero) le semionde negative. Quindi un circuito semplice semplice (un integratore, tra un po' vedrete di che si tratta) esegue una sorta di media temporale attraverso integrazione. Dunque non

viene affatto calcolato il valore medio del quadrato della grandezza!

Per chiarire cosa questa operazione comporti, supponiamo di avere $V(t) = V_0 \cos(\omega t)$ (sappiamo già che $V_{rms} = V_0/\sqrt{2}$, in questo caso). Immaginiamo ora di tagliare la parte negativa e di fare la media temporale della sola semionda positiva, che chiamo $V_+(t)$:

$$\langle V_+ \rangle = \frac{1}{T} \int_{-T/2}^{T/2} V_+(t) dt = \frac{1}{T} \int_{-T/4}^{T/4} V_+(t) dt , \quad (11)$$

dove ho cambiato gli estremi di integrazione perché tanto, negli intervalli $(-T/2, -T/4]$ e $[T/4, T/2)$ la funzione fa zero (l'ho tagliata!). L'integrale da fare è del tipo:

$\int_{-T/4}^{T/4} \cos(\omega t) dt$, il cui risultato è $2/\omega = T/\pi$ (provate!). C'è poi da moltiplicare questo integrale per $1/T$, da rimettere in situ l'ampiezza V_0 , e da moltiplicare il risultato per 2 allo scopo di tenere conto del fatto che la media è stata misurata solo sulle semionde positive (e si suppone che quelle negative contino allo stesso modo): alla fine si ottiene, con ovvio significato dei termini, $\langle V_\pm \rangle = 2V_0/\pi$. Il fattore $2/\pi$ così ottenuto *non* è uguale al fattore di proporzionalità $1/\sqrt{2}$, quello corretto per la valutazione di V_{rms} nel caso sinusoidale. In particolare, esso è inferiore di quanto dovuto per un fattore $\pi/(2\sqrt{2}) \approx 1.1$.

Il costruttore tiene conto di questo nella calibrazione dello strumento. Per esempio, nel multmetro analogico le scale per le grandezze alternate (marcate in rosso sul quadrante) non sono "allineate" con quelle per le grandezze continue (fateci caso quando vi capita!). Per il multmetro digitale, invece, è probabile che il costruttore

tenga conto di questo aspetto inserendo opportune informazioni nel software che gestisce la visualizzazione sul display.

La conseguenza pratica ovvia, però, è che la calibrazione vale, nei limiti dichiarati dal costruttore, *solo per segnali sinusoidali*. In altre parole, usare il multmetro per determinare il valore rms di segnali che siano non sinusoidali comporta un errore di calibrazione, che dovrebbe risultare nella sovrastima per un fattore ≈ 1.1 . È possibile che possiate accorgervi di ciò confrontando il valore rms (presunto) fornito dal multmetro con quello determinato a partire dalle misure con l'oscilloscopio, in particolare usando una forma d'onda quadra.

Ci sono poi ulteriori aspetti critici: come vedremo, l'operazione di integrazione dipende dalla frequenza. Questo è il motivo per cui il produttore del multmetro stabilisce un range di frequenze in cui la lettura è corretta (cioè affidabile entro la precisione dichiarata, che è dell'ordine dello 0.8-1%). Per il multmetro digitale tale range è dichiarato 40-400 Hz, ma sperimentalmente si dimostra che ci si può spingere oltre, fino ad alcuni kHz. Infine, e anche questo lo vedremo per bene, di fatto il taglio delle semionde negative, essendo affidato a un diodo a giunzione, viene eseguito solo quando la tensione sale al di sopra di un certo valore di soglia, tipicamente dell'ordine di poche centinaia di mV: ciò rende ancora meno affidabile la lettura rms di grandezze con piccola ampiezza. Se nel multmetro digitale il software del display può, almeno in parte, limitare gli effetti di questo problema, nello strumento analogico si è costretti a usare scale *non lineari* (tacchette non equispaziate tra loro): fateci caso quando vi capiterà di avere sotto mano il multmetro analogico!

Digitalizzazione (e campionamento): una breve introduzione

francesco.fuso@unipi.it

(Dated: version 1b - FF, 8 novembre 2018)

È linguaggio comune distinguere le grandezze misurate e i metodi di misura tra *analogici* e *digitali*. Ormai da alcuni decenni il mondo viaggia veloce verso la diffusione capillare di metodi digitali pressoché per tutte le necessità pratiche, in particolare quelle legate allo scambio e al trattamento delle informazioni. Nel nostro piccolissimo, anche noi possiamo toccare con mano vantaggi (molti) e svantaggi (pochi, legati soprattutto alla scelta dell'hardware) dell'approccio digitale, grazie in particolare all'uso di Arduino. Questa nota intende fornire un minimo di background utile per affrontare questo approccio.

I. ANALOGICO VS DIGITALE

Le grandezze di interesse per l'elettricità e l'elettronica hanno generalmente valori che "appaiono" continui, cioè non discreti, né quantizzati [1]. La misura di grandezze di questo tipo si dice talvolta *analogica*. Un buon esempio di misura analogica è quella che si fa con uno strumento a lancetta, dove idealmente la lancetta può muoversi assumendo con continuità qualsiasi posizione angolare e la lettura dello spostamento angolare può idealmente, cioè trascurando limitazioni tecniche ed eventualmente fondamentali, essere amplificata a volontà.

Ci sono poi delle grandezze la cui misura si effettua tramite un valore rappresentato da un *numero intero*, operazione ovvia per grandezze discrete, per esempio gli atomi contenuti in un campione, i fotoni emessi per unità di tempo da una sorgente, gli elettroni presenti in un pezzo di materiale, e così via. Quando grandezze continue sono misurate da numeri interi la misura si dice talvolta *digitale*. Convenzionalmente chiameremo *digit* l'unità arbitraria di digitalizzazione, adimensionale, e supponiamo generalmente possibile convertire la misura digitale in unità fisiche attraverso un qualche fattore di calibrazione.

La misura analogica permette di ottenere un valore idealmente accurato, cioè con tante cifre significative quante sono consentite dalla sensibilità, o accuratezza, dello strumento. La misura digitale fornisce un valore intero la cui accuratezza è, nella migliore delle ipotesi, pari al singolo digit. Esistono numerose considerazioni che possono spingere a favore dell'una o dell'altra tipologia di misura. Qui facciamo riferimento alla circostanza che, se vogliamo eseguire misure automatizzate usando un computer che ha una natura inerentemente digitale (vive di uni e zeri, cioè di numeri interi), siamo obbligati a convertire grandezze continue, cioè, forzando la nomenclatura, analogiche, in grandezze discrete, cioè digitali.

A. Digitalizzatore modello

Grazie alla vasta diffusione dei digitalizzatori, motivata soprattutto dalla diffusione della consumer electronics, sono state messe a punto diverse tecnologie specifiche, molto efficaci e spesso complicate da descrivere. Tutta-

via è possibile identificare un semplice meccanismo base in grado di spiegare come sia possibile digitalizzare un segnale analogico. In sostanza, quello che qui descriviamo è il principio di funzionamento di un *convertitore analogico-digitale* (ADC - Analog to Digital Converter, o convertitore A/D) modello, che può essere considerato il cuore di ogni dispositivo che permetta misure digitali.

Gli ingredienti fondamentali del modello sono:

1. un *clock*, ovvero un dispositivo in grado di generare impulsi, di durata virtualmente trascurabile, equispaziati nel tempo, così come fa un orologio;
2. un *contatore*, cioè un dispositivo in grado di contare gli impulsi di cui sopra, dotato di un circuito di reset (di azzeramento) opportunamente comandabile;
3. un'*interfaccia digitale*, cioè, per intenderci, un processore in grado di trattare (registrare, visualizzare, etc.) il conteggio prodotto dal contatore creando un "uscita digitale";
4. un *generatore di rampa*, cioè un dispositivo in grado di fornire una d.d.p. crescente linearmente con il tempo e dotato di un *trigger*, cioè tale che la partenza della rampa sia comandabile dall'esterno [2];
5. un *comparatore*, cioè un dispositivo in grado di comparare i livelli (le d.d.p., riferite ovviamente alla stessa linea di massa) di due ingressi e fornire in uscita un segnale digitale dipendente dall'esito della comparazione (per intenderci, uno o zero a seconda che prevalga uno o l'altro dei due segnali in ingresso).

La Fig. 1 illustra lo schema a blocchi semplificato del sistema [pannello (a)] e mostra uno schemino della temporizzazione (timing) del circuito, ovvero della sequenza temporale delle operazioni da esso compiute. Prima di andare avanti, osservate che le grandezze considerate sono d.d.p. (tanto quella incognita da misurare che quella della rampa), per cui un digitalizzatore fatto secondo questo modello è fondamentalmente un voltmetro (digitale). Inoltre notate che il conteggio degli impulsi

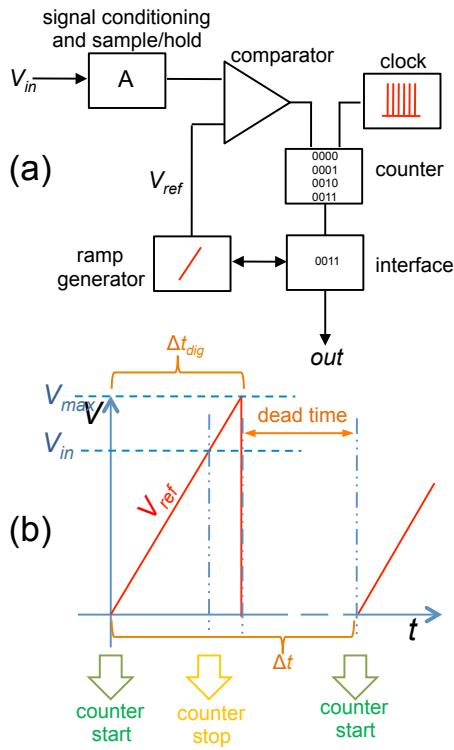


Figura 1. Schema a blocchi semplificato (a) e schema del timing di funzionamento (b) di un digitalizzatore modello; V_{in} e V_{ref} indicano rispettivamente la d.d.p. del segnale (incognita da misurare) e la d.d.p. prodotta dal generatore di rampa che aumenta linearmente nel tempo. Lo schema a blocchi non riporta esplicitamente, per semplicità tipografica, lo stadio di sample-and-hold brevemente citato nel testo.

di clock eseguito dal contatore è un’operazione inerentemente digitale, dato che il suo esito è sicuramente un numero intero. Dunque se si riesce a “collegare” la misura del tempo con quella della d.d.p. in ingresso si è sulla buona strada per costruire un digitalizzatore.

Agli ingressi del comparatore sono inviati la d.d.p. da misurare, indicata in figura con V_{in} , e la d.d.p. V_{ref} prodotta dal generatore di rampa. Supponendo che la partenza della rampa, cioè il trigger del generatore, sia sincrona con il reset del contatore, il contatore segna zero all’inizio del processo, quando $V_{ref} = 0$. Inoltre supponiamo che l’uscita del comparatore sia in grado di “bloccare” il conteggio e comandare le altre operazioni necessarie per fornire l’uscita digitale nell’istante in cui essa cambia di livello. Il contatore conta (1, 2, 3,...) fin quando $V_{in} < V_{ref}$; non appena $V_{in} \geq V_{ref}$ l’uscita del comparatore cambia il proprio livello, il processo si interrompe e l’interfaccia digitale acquisisce (“blocca”) il conteggio e inizia a elaborare l’uscita digitale. Grazie alla linearità della rampa con il tempo, questo conteggio è proporzionale al valore del segnale in ingresso, realizzando di fatto un “collegamento” lineare (di proporzionalità diretta) tra grandezza analogica da misurare e uscita in formato digitale. Un’opportuna calibrazione consente di ricondurre

il conteggio al valore in unità fisiche di V_{in} entro una determinata incertezza di calibrazione.

Inoltre è ovvio che, nelle implementazioni pratiche, il segnale incognito viene “condizionato”, cioè opportunamente amplificato o attenuato, eventualmente cambiato di segno e traslato rispetto allo zero, prima di arrivare all’ingresso del comparatore. In assenza di questo condizionamento non sarebbe possibile, per esempio, misurare con il nostro digitalizzatore modello dei segnali negativi. Anche il circuito di condizionamento contribuisce all’incertezza di calibrazione.

Infine è evidente che tutte le operazioni che abbiamo supposto istantanee (partenza della rampa a seguito del trigger, cambio del livello in uscita dal comparatore, etc.) richiedono in realtà un tempo finito che potrebbe anche costituire un’ulteriore sorgente di incertezza.

In ogni caso, poiché la conversione da analogico a digitale è in pratica una misura di tempo, è ovvio che essa richiede del tempo per essere completata. Occorre infatti un tempo minimo, che qui chiamiamo Δt_{dig} e facciamo corrispondere al tempo necessario perché la rampa passi dal livello zero al livello massimo che essa può assumere (indicato von V_{max} in figura), affinché la digitalizzazione abbia luogo. Inoltre l’uscita digitale viene prodotta e resa disponibile dopo un ulteriore intervallo di tempo, che di fatto è un tempo morto nei confronti della digitalizzazione e che può risentire di fluttuazioni e latenze nelle operazioni digitali necessarie per elaborare l’uscita digitale. Qui chiamiamo Δt la somma di Δt_{dig} con tale tempo morto: nella pratica è proprio Δt a determinare la capacità del digitalizzatore di “seguire” (o registrare) l’evoluzione temporale di una d.d.p. che varia nel tempo, operazione frequentissima dal punto di vista pratico.

Seguire l’evoluzione temporale di un segnale facendone misure digitali in istanti necessariamente discreti (nella nostra terminologia questi istanti hanno la durata finita Δt_{dig} e si ripetono con periodo Δt) significa effettuare un’operazione di *campionamento* temporale del segnale: infatti si costruisce un campione di misure della stessa grandezza effettuate a tempi successivi. Quindi digitalizzazione e campionamento sono concetti inevitabilmente connessi tra loro, ma in questa nota ci interessiamo soprattutto alla digitalizzazione, rimandando ulteriori considerazioni sul campionamento. Inoltre, specialmente nel caso di segnali rapidamente variabili, è necessario che essi vengano mantenuti costanti all’ingresso del comparatore almeno per il tempo Δt_{dig} . A questo scopo occorre un’altra operazione di condizionamento del segnale che viene normalmente attuata da un circuito detto *sample-and-hold*. Un semplice modello di sample-and-hold è un condensatore che viene caricato in maniera praticamente istantanea mantenendosi a un potenziale costante e pari a V_{in} per la durata della digitalizzazione. Al termine, cioè durante il tempo morto, il condensatore deve essere scaricato per permettere una nuova digitalizzazione, e quindi occorre un opportuno circuito di controllo in grado di compiere tali operazioni in maniera sincrona con l’operazione del digitalizzatore [3].

La descrizione che abbiamo presentato mette in evidenza quali siano le due principali figure di merito di un digitalizzatore:

- il massimo rate di campionamento, che è evidentemente dipendente dalla rapidità di risposta del comparatore e dalla ripidità della rampa prodotta; esso è normalmente espresso in Sa/s (samples per secondo) e qualche volta in Hz (eventi di campionamento per secondo); nella nostra descrizione esso corrisponde a $1/\Delta t$;
- la dinamica, o profondità di digitalizzazione, cioè il massimo numero di conteggi che può essere registrato, normalmente espresso in bit dato che nel mondo digitale vige la codifica binaria; tale dinamica è evidentemente legata, nel nostro modello, sia alla ripidità della rampa che al periodo del clock.

Entrambe queste figure di merito hanno ripercussioni sulla qualità della misura. Il massimo rate di campionamento influenza direttamente la possibilità di seguire le variazioni di segnali rapidamente dipendenti dal tempo. La dinamica, o profondità di digitalizzazione, ha a che fare con la sensibilità della misura. Infatti la digitalizzazione può essere interpretata come un processo di *binning*, in cui il segnale analogico va a riempire un certo numero di bins che dipende dal valore del segnale stesso. L’ampiezza di ogni bin, che rappresenta la sensibilità della misura in unità fisiche, è evidentemente determinata dal rapporto tra V_{max} (che è il fondo scala, o portata, della misura) e numero di bins.

Inoltre è evidente che, a prescindere da ogni altra eventuale causa di errore, esiste un’*incertezza di digitalizzazione* che vale (almeno) un digit e che possiamo interpretare come di natura prevalentemente statistica. Infatti la capacità dinamica finita implica che il comparatore scatti quando il segnale si trova all’interno di un intervallo la cui ampiezza è pari a quella di un singolo bin. È prassi comune, e motivata da considerazioni tecniche e concettuali, largheggiare nella definizione di questa incertezza: nella misura essa diventa di fatto una barra di errore, e in genere si considera una (semi-)barra di errore di almeno un digit, o più nel caso in cui la digitalizzazione comporti diversi processi “a cascata” [4].

Oltre a sampling rate e profondità di digitalizzazione, altre figure di merito che possono essere dedotte dalla descrizione del modello sono la *linearità* e la presenza di *offset*. La prima di queste caratteristiche ha a che fare con la linearità della rampa che produce V_{ref} , la seconda, oltre che da altre cause tecniche, con l’eventuale presenza di un termine non nullo in V_{ref} all’istante iniziale della rampa stessa.

II. CARATTERISTICHE PRINCIPALI DI ARDUINO

Arduino (nella versione Uno) costituisce per noi un paradigma di *interfaccia A/D e D/A*, cioè di interfaccia

tra mondo analogico (un esperimento e le sue misure) e digitale (il computer) in entrambi i versi. Dal punto di vista pratico Arduino è una schedina connessa da un lato all’esperimento tramite fili e boccole e dall’altro al computer tramite porta USB. Il termine *paradigma* significa che, pur essendo nato per scopi prevalentemente ludico/didattici, noi faremo finta che Arduino sia uno strumento adatto a consentire acquisizione dati e controllo automatizzato dei nostri semplicissimi esperimenti a prescindere da eventuali carenze hardware.

Il cuore di Arduino è un microcontroller (modello ATMEL ATmega328 per la scheda Arduino Uno) dotato, tra le altre funzioni, di 8 digitalizzatori virtualmente indipendenti collegati alle porte denominate A0 - A7. Essi possono campionare segnali (d.d.p.) analogici e convertirli in interi con una dinamica di 10 bit, corrispondenti a $2^{10} = 1024$ livelli distinti corrispondenti al range 0 – 1023 digit. Salvo le ulteriori precisazioni che discuteremo in seguito, questi livelli corrispondono in unità fisiche all’intervallo di d.d.p. tra (circa) zero e un valore massimo che è tipicamente $V_{max} \simeq 5$ V, ottenuto direttamente dalla tensione di alimentazione che proviene dalla porta USB del computer. Dunque la sensibilità della misura è di circa $5/1023 \sim 5$ mV (e la portata è pari a $\simeq 5$ V). Esiste inoltre la possibilità di operare con una tensione di riferimento generata internamente alla scheda Arduino, che comporta un valore nominale $V_{max} = 1.1$ V e dunque una sensibilità $1.1/1023 \sim 1$ mV (con portata nominale 1.1 V).

Le informazioni che riguardano il tempo di digitalizzazione e il rate di campionamento non sono riportate in maniera chiara nei datasheets. Nella configurazione (“overclocked”) con cui normalmente impieghiamo Arduino si ha tipicamente $\Delta t_{dig} \simeq 12 - 15 \mu\text{s}$ (potrà essere stimato in alcune esercitazioni), con variazioni dipendenti da fluttuazioni e dall’impiego di diverse schede; l’esperienza pratica dimostra che, al minimo, $\Delta t \sim 40 - 50 \mu\text{s}$, corrispondente a un rate di campionamento massimo di circa 20 kSa/s. Per i nostri scopi i digitalizzatori di Arduino possono essere considerati in prima approssimazione come lineari con offset prossimo allo zero (queste caratteristiche potranno essere verificate sperimentalmente in una specifica esercitazione).

Dal punto di vista concettuale il microcontroller di cui è dotato Arduino riproduce, in forma ridotta e parziale, il processore (CPU) di un qualsiasi computer e quindi come questo ha bisogno in primo luogo di essere istruito sulle operazioni che vogliamo che compia. In un normale computer questo è, grosso modo e senza entrare nei dettagli, quanto viene eseguito dalla combinazione di programma e sistema operativo. Una versione molto semplificata di sistema operativo specifico è residente in una memoria permanente contenuta nel microcontroller.

Le istruzioni di programma possono essere date scrivendo un semplice testo, detto *sketch*, all’interno di un ambiente interattivo, detto IDE, specifico, cioè un programma che si chiama Arduino, Arduino IDE, o Arduino Programming e che è rilasciato per tutti i principali

sistemi operativi. Questo ambiente interattivo è presente nei computer del laboratorio e si individua facilmente, essendo identificato dall'icona di Arduino (un infinito su sfondo verde acqua). Lo sketch, che è composto di parti distinte, tutte funzionali e necessarie, è scritto con una sintassi che ricorda molto da vicino quella del linguaggio C (più precisamente si tratta di una sorta di sottoinsieme del C, implementato nel sottoinsieme di un ambiente/linguaggio che si chiama Processing). Naturalmente lo sketch contiene anche delle istruzioni specifiche per controllare Arduino, per esempio quelle che stabiliscono se le varie porte disponibili devono essere considerate come ingressi o uscite, quelle che eseguono la lettura di una porta analogica o la “scrittura” (determinazione di livello “alto” o “basso” [5]) per una porta digitale. Fortunatamente, la sintassi di queste istruzioni è abbastanza ben comprendibile e in molti casi addirittura auto-esplicativa.

Il programma Arduino IDE presente sul computer provvede, una volta terminata la redazione ed eseguiti con successo alcuni test preliminari sulla sintassi, a fare l'*upload*, cioè trasferire con un apposito comando (l'icona è una freccina) il contenuto dello sketch, debitamente compilato, in una memoria non volatile (di tipo “flash”) di cui è dotato il microcontrollore. Il trasferimento avviene sfruttando la comunicazione seriale emulata su USB tra computer e scheda Arduino. Purtroppo le dimensioni della memoria non volatile sono piccolissime (32 kB), per cui lo sketch deve necessariamente essere semplice e breve e le variabili previste (tipicamente array) devono avere piccole dimensioni. Una volta che lo sketch compilato è stato trasferito, le istruzioni diventano *residenti* nel microcontrollore, che dunque le eseguirà finché non verrà in qualche modo resettato, per esempio sovrascrivendo lo sketch con uno nuovo (non basta scollegare la scheda Arduino alla presa USB per cancellare il programma). Normalmente nella fase di trasferimento del programma dal computer ad Arduino alcuni led presenti sulla scheda si accendono e si spengono a mostrare che c'è una comunicazione in corso.

Nelle nostre esercitazioni la comunicazione seriale emulata su USB viene anche impiegata per “controllare” Arduino, cioè, per esempio, per avviare l'acquisizione dei dati e impostarne eventuali parametri, oltre che per raccogliere i dati stessi. A questo scopo si fa uso di semplici script di Python che, come tantissimi altri linguaggi o ambienti, ha la possibilità di inviare e ricevere istruzioni via porta seriale emulata su USB. Il vantaggio pratico è quello di integrare in un unico script le funzioni di controllo con la lettura dei dati e la registrazione dei files. Notate che la libreria che Python usa per gestire la comunicazione seriale, detta **serial**, può allo stato attuale essere utilizzata solo con Python 2.x, cioè lanciando lo script da terminale.

Poiché la comunicazione seriale è relativamente lenta (normalmente poche decine di kbaud, dove 1 baud = 1 bit/s) non è consigliabile trasferire sequenzialmente i dati da Arduino al computer durante la loro acquisizione.

Conviene piuttosto che i dati corrispondenti a un'acquisizione completa, che qui chiamiamo *record*, siano temporaneamente immagazzinati in una memoria e da qui, al termine dell'esperimento, trasferiti con tutta calma al computer. Prima di essere trasferito al computer il record viene registrato all'interno del microcontrollore, che per questo scopo sfrutta una piccolissima sezione di memoria (2 kB, di tipo SRAM), le cui dimensioni sono evidentemente tali da permettere di registrare pochi dati (tipicamente 256 coppie di dati).

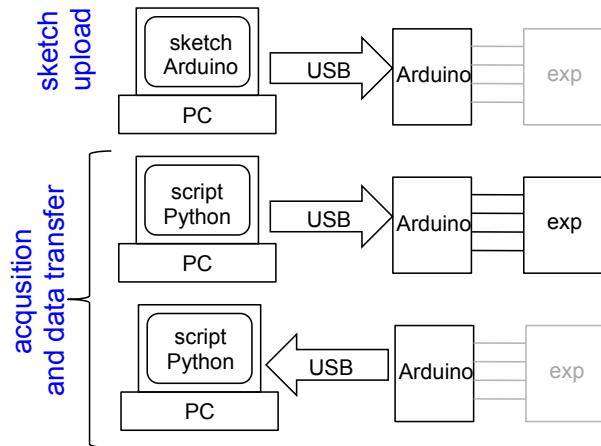


Figura 2. Diagramma logico di massima di una tipica esperienza con uso di Arduino.

Il diagramma logico che sta alla base dell'impiego di Arduino nelle nostre esercitazioni è mostrato molto schematicamente in Fig. 2:

1. si scrive uno sketch nell'ambiente Arduino (IDE, o Programming) che contiene, in un formato opportuno e con una sintassi simil-C, le istruzioni da impartire al microcontrollore (naturalmente questo sketch lo troverete già presente nei computer di laboratorio, ma potrete eventualmente modificarlo e migliorarlo, cambiandogli nome per evitare confusione);
2. lo sketch viene trasferito (*upload*) a Arduino tramite comunicazione seriale USB e caricato nella memoria non volatile del microcontrollore (una sola volta);
3. si scrive uno script di Python che serve per controllare Arduino e per permettere il trasferimento e la lettura dei dati acquisiti (anche in questo caso lo script lo troverete già nei computer e siete liberi di migliorarlo e modificarlo secondo necessità, cambiandogli nome);
4. si lancia lo script di Python, naturalmente dopo aver collegato in modo corretto i componenti necessari per l'esperimento, si aspetta un po' e, quando segnalato dalla console di Python, si sa che le

misure sono state acquisite e trasferite dalla memoria SRAM di Arduino in un file registrato nel computer;

5. il file è allora pronto per essere riaperto con un altro script di Python da fare ex-novo, che permetterà l'analisi del campione di dati (grafici, fit, etc.); tutta l'operazione di misura può essere eseguita più volte anche in modo automatico, per esempio per consentire l'acquisizione automatizzata di record più lunghi di quanto consentito dalle dimensioni della memoria SRAM.

Dato che qui siamo a riassumere le principali caratteristiche di Arduino, osserviamo che la resistenza di ingresso del digitalizzatore, come dichiarata nei datasheets, è molto alta (100 Mohm nominali) per cui i digitalizzatori di Arduino si comportano da voltmetri pressoché ideali. Oltre agli ingressi analogici, Arduino ha poi la possibilità di controllare ben 14 porte digitali configurabili come input o output, che possono quindi leggere o produrre livelli alti e bassi (1 e 0 binari). Configurate come uscite queste porte sono utili per “controllare” l'esperimento, come per esempio nella carica/scarica del condensatore. La massima corrente che può essere ottenuta è, nominalmente, di 20 mA (50 mA in totale se se ne usa più di una).

Un aspetto interessante è che la d.d.p. prodotta dalle porte digitali di Arduino approssima (in genere molto be-

ne) il valore massimo V_{max} della d.d.p. che Arduino può digitalizzare. Questa circostanza permette, assumendo linearità e mancanza di offset, di determinare il fattore di calibrazione dei digitalizzatori dividendo il valore della d.d.p. in uscita da una porta digitale per 1023 (il numero di livelli digitalizzati, comprendendo anche il livello 0). Chiameremo *alternativa* questa semplice procedura di calibrazione per differenziarla da quella, ordinaria, eseguita per confronto con uno strumento calibrato.

Sei tra le porte digitali, quelle marcate con un tilde nelle serigrafie, possono operare come output in una modalità detta **PWM** (*Pulse-Width Modulation*). In questa modalità le porte generano treni di impulsi a frequenza fissa (e piuttosto bassa, cioè minore di 1 kHz) con duty-cycle regolabile via software su 8 bit, cioè 256 distinti valori [6]. Anche la disponibilità di questa tipologia di segnale può essere utile negli esperimenti, dato che, come vedremo, integrando temporalmente i treni di impulsi si ottiene una d.d.p. quasi continua il cui livello è proporzionale al duty-cycle.

Infine, avrete sicuramente modo di usare alcune di queste porte come ingressi digitali soprattutto in esperimenti in cui è richiesta un'acquisizione dati *sincrona*, cioè triggerata da un segnale esterno. Più in generale, grazie alla sua versatilità Arduino può essere impiegato in diverse configurazioni sperimentali, come descriveremo in future note.

[1] Naturalmente ci sono importantissimi controesempi a questa affermazione dovuti in particolare alla natura discreta della carica elettrica. Provate per esempio a trovare il legame tra d.d.p. e carica accumulata in un condensatore di capacità piccolissima, dell'ordine dell'aF. Tuttavia nel mondo macroscopico il carattere discreto delle grandezze di interesse pratico non può essere apprezzato, almeno non nelle ordinarie misure di nostro interesse.

[2] Un circuito di questo tipo, cioè un generatore di rampa triggerabile, è anche impiegato negli oscilloscopi, dove serve per generare la *sweep* (o spazzata temporale).

[3] È evidente che il sample-and-hold e tutto l'ambaradan necessario al suo funzionamento sono componenti cruciali di un digitalizzatore, dovendo operare su scale temporali particolarmente brevi. Probabilmente questi componenti sono i principali responsabili del comportamento “erroneo” che si riscontra quando Arduino viene impiegato per digitalizzare segnali variabili nel tempo (scalettature e andamenti

irregolari visibili in alcune esercitazioni pratiche).

[4] L'incertezza di digitalizzazione corrisponde a una sorta di errore di lettura e quindi le attribuiamo un carattere prevalentemente statistico. La circostanza che l'incertezza di digitalizzazione possa eccedere il singolo digit è evidente leggendo il manuale del multmetro digitale, dove, per qualche portata, essa vale anche alcuni digit.

[5] Arduino segue una convenzione molto diffusa, detta **TTL** (Transistor-Transistor Logic) per stabilire i livelli alto e basso, che corrispondono agli uni e zeri del mondo digitale. In termini pratici, livello alto significa una d.d.p. di circa 5 V (per convenzione compresa tra 2.4 e 5 V) rispetto alla linea di terra, o massa, livello basso significa una d.d.p. nulla (per convenzione minore di 0.4 V).

[6] Il duty-cycle rappresenta, in valore relativo o percentuale, la quantità di tempo in un singolo ciclo in cui il segnale si trova allo stato alto. Un segnale (periodico) con duty-cycle del 50%, o di 0.5, in uscita da una porta PWM di Arduino rappresenta di fatto un'onda quadra *simmetrica*.

Arduino e campionamento/digitalizzazione di segnali continui

francesco.fuso@unipi.it

(Dated: version 6b - FF, 8 novembre 2018)

Queste nota ha lo scopo principale di illustrare le modalità di funzionamento di base tipiche delle esercitazioni pratiche in cui si fa uso di Arduino [1], prendendo spunto da quella finalizzata a esaminarne il comportamento in misure di d.d.p. continue e determinarne la calibrazione. Essendo la prima occasione di impiego di Arduino, adeguato spazio è dato a illustrare alcuni dettagli pratici.

I. INTRODUZIONE

L'esperienza considerata rappresenta una semplicissima ("la più semplice") applicazione di acquisizione automatizzata di dati. In buona sostanza c'è una d.d.p. costante, ovvero supposta tale, prodotta da un partitore di tensione collegato al solito generatore di d.d.p. in uso in laboratorio. Questa d.d.p. deve essere digitalizzata e acquisita un gran numero di volte in istanti successivi allo scopo di creare un *campione* disponibile per analisi statistiche (calcolo della media, della deviazione standard, histogrammi delle occorrenze, etc.), oltre a permettere di eseguire una calibrazione del digitalizzatore tramite misura di diversi valori e confronto con la lettura del multimetro digitale, usato qui come riferimento.

Per ottenere gli scopi della presente esercitazione pratica è necessario istruire Arduino a compiere una sequenza di misure della d.d.p., che va collegata a una delle porte analogiche di cui è dotato (nell'esempio è la porta collegata al pin A0). Queste misure produrranno in modo automatico i campioni di nostro interesse, registrati in un file di due colonne (tempo e valore digitalizzato) disponibile per le ulteriori analisi.

Poiché, come chiariremo nel seguito, il trasferimento dei dati da Arduino al computer è generalmente lento (richiede secondi) e visto che Arduino dispone di una (piccola) memoria interna, l'istruzione impartita ad Arduino prevede che esso immagazzini temporaneamente i risultati delle misure nella sua memoria interna, per poi trasferire al computer il *record* contenente tutti i dati, in "un colpo solo" al termine dell'acquisizione.

II. CONFIGURAZIONE DI MISURA

L'esercitazione pratica prevede di eseguire in maniera automatica molte misure (centinaia o migliaia) della stessa d.d.p., qui chiamata ΔV . Tale grandezza viene *digitalizzata* da Arduino: dunque, a meno di non eseguire una *calibrazione*, della quale ci occuperemo in seguito, essa sarà data da un *numero intero* (misurato in *unità arbitrarie di digitalizzazione*, che qui chiamiamo anche *digit*) necessariamente compreso tra 0 e 1023. Infatti la dinamica, o profondità di digitalizzazione, di Arduino è 10 bit, ovvero i livelli di digitalizzazione possibili sono $2^{10} = 1024$.

Le tante misure vengono acquisite in successione, dunque a istanti diversi. In questa esperienza l'acquisizione è *asincrona*, cioè non deve partire in contemporanea con qualche ben definito evento esterno [2], e *non interessa* conoscere l'istante in cui avviene l'acquisizione, poiché non abbiamo necessità di studiare l'andamento temporale della grandezza considerata, che è supposta continua. Tuttavia, come preparazione a ulteriori esperienze con Arduino e anche allo scopo di studiare l'incertezza nella misura dei tempi, Arduino è predisposto per acquisire il *time stamp*, cioè l'indicazione dell'"istante" di digitalizzazione (riferito naturalmente a un tempo zero opportunamente definito). In linea di massima, gli istanti di digitalizzazione delle singole misure potrebbero essere scelti arbitrariamente, però, come sarà evidente nel seguito, è estremamente più semplice impostare l'esperimento in modo che essi siano *nominalmente* equispaziati per un tempo Δt : l'analisi dei dati corrispondenti permetterà di verificare entro quale accuratezza tale equispaziatura sia effettivamente realizzata.

Facciamo subito due osservazioni molto importanti, anche se non necessariamente rilevanti per la presente esperienza:

1. gli intervalli di tempo tra una misura e la successiva sono gestiti, cioè, di fatto, decisi, dal programma che gira nel microcontrollore. Di conseguenza essi sono affetti da un'incertezza che, in generale, può essere non trascurabile [3];
2. la digitalizzazione non può essere istantanea e il campionamento avviene in un breve intervallo di tempo, altrove chiamato Δt_{dig} , che è avviato dalle istruzioni del programma che gira nel microcontrollore, ma che normalmente ha luogo in una frazione del tempo Δt che intercorre fra due digitalizzazioni successive.

A. Partitore con potenziometro

Poiché è di interesse misurare d.d.p. di diverso valore, e tenendo conto che in laboratorio non è disponibile un generatore di tensione variabile, è ovvio che la d.d.p. da misurare venga prodotta da un *partitore di tensione* collegato al solito alimentatore che siete ormai abituati ad usare. Per evitare di essere vincolati a valori prefissati del rapporto di partizione, come si verifica quando si ha

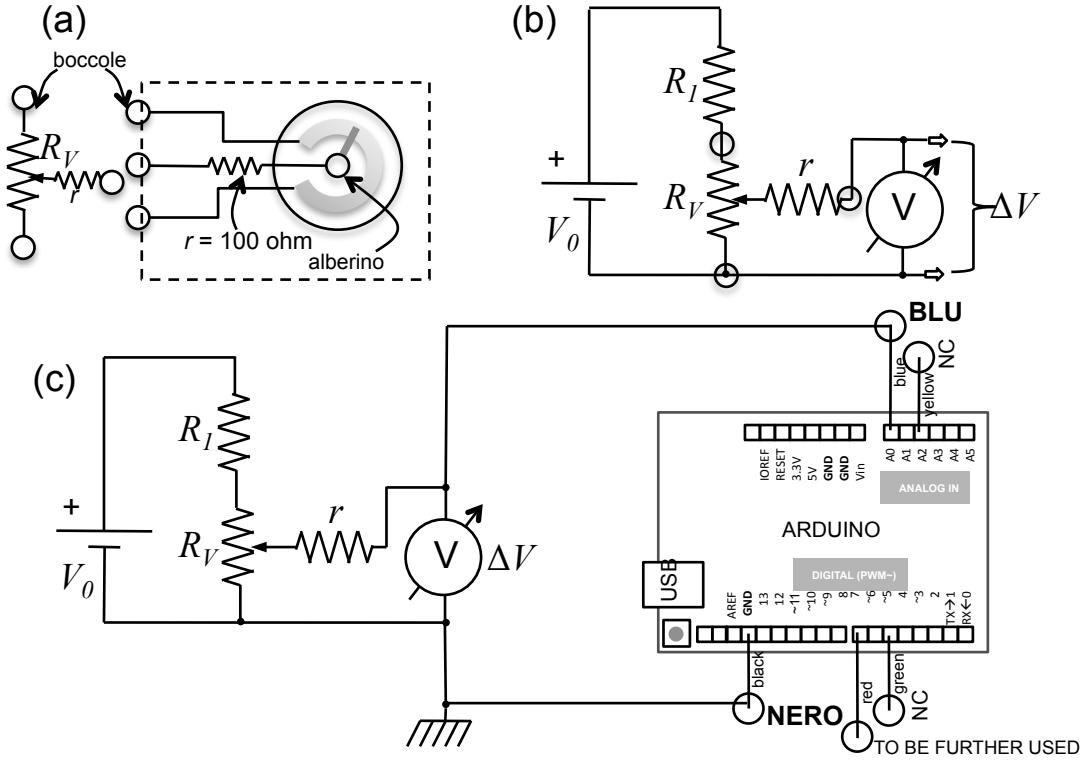


Figura 1. Rappresentazione schematica e costruttiva del potenziometro “visto da sotto” (a), schema del partitore di tensione (b), configurazione del circuito di misura comprendente Arduino (c). Nel pannello (a) è riportato uno dei simboli con cui si indica un potenziometro negli schemi elettronici. Nel pannello (c) è rappresentata una visione molto schematica e non in scala della scheda Arduino Uno rev. 3 SMD edition usata nell’esperienza. Ci sono cinque collegamenti a altrettanti pin della scheda che terminano con boccole volanti di diverso colore, secondo quanto indicato in figura: solo due boccole devono essere collegate (NC significa non collegato). Notate che la boccoola rossa collegata al pin 7 potrà eventualmente essere impiegata secondo quanto descritto in seguito. La “spazzolina” sulla linea di circuito che va al pin GND indica un collegamento a massa, ovvero a terra (il collegamento a terra è realizzato attraverso l’alimentatore del PC di laboratorio).

a disposizione un numero limitato di resistori, per questa esperienza il partitore di tensione include un resistore variabile, o *potenziometro*.

Il potenziometro è un dispositivo elettro-mecanico che, almeno grossolanamente, può essere visto come un contatto strisciante mobile su una pista di materiale conduttore dotato di alta resistività. Il contatto strisciante è solidale a un alberino, la cui rotazione, quindi, fa assumere una diversa resistenza tra il contatto strisciante stesso e le due estremità della pista conduttrice. La Fig. 1(a) il-

lustra schematicamente la realizzazione e riporta un simbolo convenzionale del potenziometro. Notate che, in genere, il potenziometro ha tre terminali, come un figura. La resistenza tra contatto strisciante (terminale “centrale”) e uno dei due estremi della pista conduttrice (uno degli altri due terminali) varia tra 0 (circa) e un valore massimo R_V in funzione della rotazione dell’alberino; nel contempo, la resistenza tra contatto strisciante e l’altro terminale varia tra R_V e (circa) 0. In laboratorio sono disponibili diversi potenziometri, la maggior parte dei quali ha $R_V = 4.7$ kohm o $R_V = 470$ kohm.

La schema del partitore di tensione è rappresentato in Fig. 1(b), dove il generatore di d.d.p. è quello disponibile in laboratorio ($V_0 \simeq 5$ V). Si vede come siano presenti altre due resistenze: R_L , da scegliere nel banco delle resistenze (nelle mie prove $R_L = 680$ ohm nominali) e $r = 100$ ohm (nominali), saldata direttamente al terminale centrale del potenziometro, dunque parte del tela-

ietto che ospita questo dispositivo. Queste resistenze sono incluse nel circuito a fini “protettivi”, cioè per evitare che nel partitore fluisca una corrente troppo alta (comporterebbe possibile bruciatura del fusibile, e anche del potenziometro, che può dissipare una potenza massima di 1 W, tipicamente). Per come è configurato il circuito, la rotazione dell’alberino del potenziometro permette di

ottenere in uscita dal partitore una d.d.p. variabile con continuità da (circa) 0 a un certo valore massimo, determinato dai valori di V_0 , R_1 , r (potete facilmente dimostrarlo con le regoline dei partitori di tensione) [4]. Questa d.d.p. può, e deve, essere misurata continuativamente: allo scopo si usa il tester digitale, che ha resistenza interna sicuramente maggiore di quella del potenziometro e dunque “perturba” in modo trascurabile il circuito. Nel mio esempio, dove ho impiegato il potenziometro con $R_V = 4.7$ kohm, ho ottenuto $\Delta V \sim 0 - 4.5$ V.

L’uscita del partitore deve essere inviata all’ingresso (porta analogica) di Arduino prescelto per la misura, che in questo esempio corrisponde al pin A0. Il pin in questione si trova, assieme ad altri, su un connettore a pettine di tipo femmina. Su di esso è innestato una maschia con dei cavetti saldati che terminano con boccole volanti: si usano colori diversi per cavetti e boccole diverse, e quello del pin A0 è il blu. Ricordate che la misura di una tensione richiede di usare due fili (è una *differenza* di potenziale): l’altro filo, che deve essere collegato alla linea connessa con il negativo dell’alimentatore, va a uno dei pin marcati con GND. L’indicazione è un’abbreviazione di ground, cioè terra: infatti questa boccola è collegata alle linee di riferimento che Arduino usa come potenziale nullo (*massa*) che, attraverso la connessione USB, sono connesse alle linee di massa del computer e di lì, tramite il connettore di alimentazione, alla terra dell’impianto di distribuzione elettrica. Boccola e filo del collegamento di massa, o terra, sono di colore nero. Ricordate anche che, per come è costruito, il digitalizzatore di Arduino accetta in ingresso solo d.d.p. *positive* (o nulle) rispetto alla linea di terra. Pertanto fate la massima attenzione a *rispettare le polarità*: la boccola di uscita del generatore di d.d.p. che deve essere collegata alla linea di terra (boccola volante nera collegata al pin GND di Arduino) è quella nera. Se sbagliate, Arduino può salutarvi e passare nello stato di big sleep eterno.

I connettori a pettine di cui sono dotate le schede Arduino in uso in laboratorio hanno anche altre connessioni: in linea di massima non dovete usare altre boccole, tranne quella rossa, collegata alla porta digitale corrispondente al pin 7, da impiegare per gli scopi di calibrazione (alternativa) che descriveremo in seguito. Le connessioni da effettuare, esclusa quella eventuale al pin 7, sono schematizzate in Fig. 1(c).

Come informazione rilevante dal punto di vista pratico, ricordate che Arduino mantiene i programmi nella sua memoria non volatile finché questi non vengono riscritti. Di conseguenza è possibile che, collegando Arduino al circuito come in Fig. 1 senza aver preventivamente fatto l’upload dello sketch di interesse, il comportamento del circuito sia erroneo. Dunque come norma generale collegate Arduino al resto del circuito *solo dopo aver effettuato l’upload dello sketch*.

Ricordiamo infatti che Arduino viene istruito a compiere determinate operazioni attraverso un semplice programma, scritto in un file di testo (con estensione .ino) da trasferire (uploadare) nel microcontroller attraverso il

programma Arduino, o Arduino IDE, presente nei computer di laboratorio. La comunicazione con il computer e la gestione dell’esperimento vengono invece controllate da Python usando appositi script.

III. LO SCRIPT DI PYTHON

Anche se la logica suggerirebbe di partire dalle istruzioni dello sketch di Arduino, iniziamo con il commento allo script di Python.

Lo script richiede di importare due pacchetti che finora non abbiamo mai usato: il pacchetto `serial`, che serve per gestire (al meglio) la comunicazione seriale USB, e il pacchetto `time`, che permette di eseguire dei cicli di attesa con un tempo controllato. Nello script compaiono infatti delle istruzioni del tipo `time.sleep(2)` che producono un’attesa di 2 s (il valore può essere ovviamente modificato, l’unità di misura è secondi), precauzionalmente necessaria per evitare di avere problemi di intasamento della comunicazione seriale. Tenete sempre presente che nei computer di laboratorio la libreria `serial` può essere caricata solo lanciando Python da terminale (non da Pyzo).

A causa delle piccole dimensioni della memoria SRAM di Arduino, solo 2 kB, il numero di misure distinte che possono essere eseguite e registrate in una singola acquisizione è limitato (sono 256 nell’esempio qui considerato). Dato che potrebbe essere utile creare un campione di misure più grande, lo script è predisposto per eseguire un loop di diverse acquisizioni, permettendone la registrazione su un unico file [5]. Questo loop è avviato dall’istruzione `for j in range (1,nacqs+1):`, con `nacqs` da definire nello script (di default pari a 1). State attenti alla particolare sintassi: in Python l’istruzione del loop termina con un “`:`” e le istruzioni che devono essere eseguite nel ciclo sono *indentate* (tabulate, per usare il linguaggio delle macchine da scrivere), cioè rientrate rispetto al margine sinistro dello script esattamente come nella definizione di funzioni. Inoltre nella parte iniziale dello script si stabilisce la directory che conterrà i dati. Di default, alla directory dove sono raccolti i dati nei computer di laboratorio si accede con `../dati_arduino/` (si intende che Python sia lanciato avendo `Home` come directory presente). È anche necessario fornire il nome del file, che dovrete stabilire secondo i vostri gusti, compresa l’estensione (consigliata) `.txt`.

Dopo aver inizializzato la porta seriale, cioè attribuito alla variabile `ard` un valore identificativo della porta USB a cui la scheda Arduino è collegata (la sintassi è peculiare e fortemente dipendente dal sistema operativo) e specificato che la comunicazione avverrà alla velocità di 9600 baud (1 baud = 1 bit/s), non molto elevata ma sicuramente adeguata agli scopi (potrebbe essere aumentata, cioè moltiplicata per un multiplo di 2), lo script scrive sulla porta seriale un determinato valore. L’istruzione corrispondente è, per esempio, `ard.write(b'5')` che significa che alla porta seriale corrispondente alla variabile

`ard` (sarebbe il nostro Arduino) viene inviato in scrittura (`.write`, sintassi in cui si riconosce bene la concatenazione attraverso il punto in uso con Python) un carattere di tipo ASCII (il `b` dell'istruzione, che a rigore non serve usando Python 2.x) costituito dal carattere '5'.

Come sarà discusso in seguito, l'invio di questo carattere ha la duplice funzione di far partire l'acquisizione da parte di Arduino e di indicargli quanto deve valere l'intervallo temporale *nominale* Δt_{nom} tra una digitalizzazione e la successiva. L'unità di misura è, in questa esperienza, $100 \mu\text{s}$, per cui il "5" significa che i dati saranno campionati con intervalli nominali di $5 \times 100 \mu\text{s} = 500 \mu\text{s}$. La scelta di questo parametro non influenza i risultati della presente esperienza, mentre invece sarà critica per altre esperienze da svolgere in futuro. Essa sarà inoltre considerata in seguito, quando tratteremo dell'analisi dell'incertezza su Δt .

Quindi lo script attende finché sulla porta seriale, continuamente monitorata, non compaiono dei dati. Arduino li rende disponibili al termine dell'acquisizione, dunque in questo modo ci si garantisce che il record venga trasferito al computer quando effettivamente pronto. Poiché la comunicazione seriale prevede lo scambio di dati uno alla volta, la porta seriale viene letta all'interno di un loop, con un indice che gira fino al numero di dati acquisiti, cioè delle misure fatte (256, nell'esempio considerato). Notate la sintassi abbastanza specifica: essa prevede di leggere una riga (coppia di dati) alla volta attraverso l'istruzione `data = ard.readline().decode()`, che con-

tiene anche l'istruzione di decodifica dei dati stessi che devono essere interpretati come numeri (interi). Ogni coppia di dati viene aggiunta al file di testo prodotto dallo script. Al termine di ogni acquisizione del ciclo viene chiusa la comunicazione seriale con Arduino attraverso l'istruzione `ard.close()` e al termine delle operazioni il file dei dati viene anche chiuso con l'istruzione `outputFile.close()`.

Nel corso di tutto il processo è prevista la scrittura sulla console (cioè sul terminale) di indicazioni di progresso. Per agevolare alcune delle operazioni previste nell'esercitazione pratica lo script si occupa anche di calcolare valore medio e deviazione standard *sperimentale* del campione di dati digitalizzati (si intende campione di 256 punti, in questo esempio). Visto che, come sarà chiarito in seguito, i dati vengono codificati da Arduino nella forma di righe (per un totale di 256, nell'esempio qui considerato) contenenti il time stamp in unità di μs , uno spazio, il valore digitalizzato (in digit), occorre un'istruzione dalla sintassi apparentemente misteriosa, `runningddp[i]=data[data.find(' '):len(data)]`, per estrarre il dato di interesse e metterlo in un array di supporto. Quindi media e deviazione standard sono calcolate su questo array usando istruzioni standard di Python e il risultato viene scritto sulla console. Alla fine di tutto compare sulla console un bell'`end`.

Lo script, debitamente commentato, è riportato qui di seguito; esso si trova nei computer di laboratorio (nella directory `/Arduini/`) e in rete sotto il nome di `ardu2016.py`.

```

import serial # libreria per gestione porta seriale (USB)
import time # libreria per temporizzazione
import numpy

nacqs = 1 # numero di acquisizioni da registrare (ognuna da 256 coppie di punti)
Directory='../../dati_arduino/' # nome directory dove salvare i file dati
FileName=(Directory+'dataXX.txt') # nomina il file dati <<< DA CAMBIARE SECONDO GUSTO
outputFile = open(FileName, "w" ) # apre file dati predisposto per scrittura

for j in range (1,nacqs+1):
    ard=serial.Serial('/dev/ttyACM0',9600) # apre la porta seriale
    # (da controllare come viene denominata, in genere /dev/ttyACM0)
    time.sleep(2) # aspetta due secondi per evitare casini
    ard.write(b'5') # scrive il carattere per l'intervallo di campionamento
                    # in unità di 100 us << DA CAMBIARE A SECONDA DEI GUSTI
                    # l'istruzione b indica che è un byte (carattere ASCII)
    time.sleep(2) # aspetta due secondi per evitare casini
    print('Start Acquisition ',j, ' of ',nacqs) # scrive sulla console (terminale)
    # loop lettura dati da seriale (256 coppie di dati: tempo in us, valore digitalizzato di d.d.p.)
    runningddp=numpy.zeros(256) # prepara il vettore per la determinazione della ddp media e std

    for i in range (0,256):
        data = ard.readline().decode() # legge il dato e lo decodifica
        if data:
            outputFile.write(data) # scrive i dati sul file
            runningddp[i]=data[data.find(' '):len(data)] # estrae le ddp e le mette nel vettore

```

```

ard.close() # chiude la comunicazione seriale con Arduino

avgddp=numpy.average(runningddp) # analizza il vettore per trovare la media
stdddp=numpy.std(runningddp) # e la deviazione standard
print('Average and exp std:', avgddp, '+/-',stdddp) # le scrive sulla console

outputFile.close() # chiude il file dei dati
print('end') # scrive sulla console che ha finito

```

IV. LO SKETCH DI ARDUINO

Lo sketch di Arduino è scritto in un linguaggio che somiglia al C. In questo linguaggio si fa uso molto spesso di pseudo-funzioni, cioè gruppi di istruzioni che non ritornano un valore numerico. A queste pseudo-funzioni si fa riferimento con l'istruzione `void {}` (le parentesi graffe comprendono le istruzioni associate alla pseudo-funzione). Notate che pressoché tutte le singole istruzioni contenute tra le parentesi graffe devono necessariamente *terminare con un punto e virgola* (fanno eccezione, per esempio, le istruzioni che avviano un loop, per le quali il punto e virgola non deve essere usato).

Nei casi semplici, a cui fortunatamente appartiene il nostro sketch, Arduino richiede che esso sia suddiviso in diverse parti, che nella nostra implementazione sono tre poste consecutivamente una dietro l'altra: (i) dichiarazione delle variabili; (ii) inizializzazione del microcontrollore; (iii) istruzioni necessarie per le specifiche operazioni previste.

Vediamo e commentiamo brevemente il contenuto di queste tre parti.

A. Dichiarazione delle variabili

La dichiarazione delle variabili e l'allocazione dello spazio di memoria relativo è necessaria (in C, non in Python, come sapete) affinché esse possano essere correttamente interpretate nel programma. Si possono definire delle variabili vere e proprie, oppure delle *costanti*, cioè delle grandezze che non verranno mai modificate dal programma. A seconda di quello che devono rappresentare, esse saranno identificate come variabili secche (scalari) o array (vettori), intere (segnate o meno, eventualmente "lunghe") o reali (eventualmente a doppia precisione).

A seconda della tipologia di definizione cambia la quantità di memoria allocata per la variabile. Vista l'esiguità dello spazio di memoria disponibile nel microcontrollore, è

sempre consigliabile definire le variabili per quello che effettivamente serve. Per esempio, un intero standard (`int` o `unsigned int`, a seconda che debba o non debba assumere valori negativi) occupa 2 byte (1 byte equivale a 8 bit), cioè due caratteri, e permette quindi di individuare $2^{8+8} = 2^{16}$ valori interi differenti; un intero `long` richiede invece 4 byte.

Nel nostro caso abbiamo sicuramente a che fare con due distinte variabili di tipo array, quelle che vanno acquisite e registrate. Esse sono l'array denominato `V`, che contiene il valore digitalizzato della d.d.p., e quello denominato `t`, che contiene il time stamp. Nel primo caso è sufficiente definire l'array come intero a singola precisione, `int`, visto che il valore digitalizzato è necessariamente compreso tra 0 e 1023. Nel secondo caso, invece, visto che il time stamp è in unità di microsecondi e che la durata complessiva dell'acquisizione può essere "lunga" su questa scala, occorre definire la variabile come `long`: infatti, per l'intero campione di 256 misure, se per esempio l'intervallo di campionamento nominale è $\Delta t_{nom} = 500 \mu s$, l'ultima misura avviene (almeno) dopo $1.28 \times 10^3 \mu s$, numero per la cui registrazione non basta un intero a singola precisione. Le due istruzioni di definizione sono rispettivamente `int V[256];` e `long t[256];`, le quali mostrano pure che gli array sono entrambi costituiti da 256 punti.

Il resto di questa sezione dello sketch, che è riportata qui nel seguito, è piuttosto auto-esplicativa: notate che `analogPin` e `digitalPin` sono le costanti intere che indicano i pin da impiegare come porte per la lettura (analogica) e per fornire in uscita un valore di d.d.p. che sarà utile per la calibrazione alternativa, cioè le porte corrispondenti ai pin A0 e 7 della scheda. La variabile intera `start` serve come *flag*: nel seguito dello sketch ci sarà un ciclo pronto a partire quando questa variabile diventerà diversa da zero, mentre la variabile intera `delays` contiene il ritardo nominale Δt_{nom} (in μs) tra una digitalizzazione e la successiva. Infine, la variabile `StartTime`, definita `long`, serve per indicare lo zero dei tempi, secondo quanto sarà chiarito in seguito.

```

// Blocco definizioni
const unsigned int analogPin=0; // Definisce la porta A0 per la lettura
const int digitalPin=7; // Definisce la porta 7 usata come output ref
int i; // Definisce la variabile intera i (contatore)
int delays; // Definisce la variabile intera delays
int V[256]; // Definisce l'array intero V

```

```
long t[256]; // Definisce l'array t
unsigned long StartTime; // Definisce la variabile StartTime
int start=0; // Definisce la variabile start (usato come flag)
```

B. Inizializzazione

Le istruzioni di inizializzazione di Arduino devono essere incluse nella pseudo-funzione chiamata `setup()`. Pertanto esse iniziano con `void setup(){}` e le istruzioni relative sono contenute tra le parentesi graffe.

L'inizializzazione richiede di aprire la porta seriale preparandola a funzionare a 9600 baud (`Serial.begin(9600);`), di pulirne per sicurezza il buffer (`Serial.flush();`), di definire di uscita la porta indicata dalla variabile `digitalPin` e di porla a livello alto [6] per gli scopi di cui tratteremo in seguito (l'istruzione è auto-esplicativa). Notate che non è necessario definire la porta analogica come input, essendo questa la configurazione di default.

Inoltre il blocco di inizializzazione contiene due linee

```
// Istruzioni di inizializzazione
void setup()
{
    Serial.begin(19200); // Inizializza la porta seriale a 19200 baud
    Serial.flush(); // Pulisce il buffer della porta seriale
    digitalWrite(digitalPin,HIGH); // Pone digitalPin a livello alto
    bitClear(ADCSRA,ADPS0); // Istruzioni necessarie per velocizzare
    bitClear(ADCSRA,ADPS2); // il rate di digitalizzazione
}
```

C. Il loop

Le istruzioni vere e proprie del programma sono inserite in una pseudo-funzione che prevede un ciclo ed è pertanto denominata `void loop(){}` ; come al solito, anche qui le istruzioni del loop sono contenute tra le parentesi graffe.

Questo ciclo inizia con un'istruzione di monitoraggio della porta seriale, necessario perché, dopo aver caricato lo sketch nel microcontroller, esso aspetti per partire di avere ricevuto via seriale (USB) la comunicazione prodotta dallo script di Python. Allo scopo provvede il comando `Serial.available()`, che ritorna un valore diverso da zero quando qualcosa si viene a trovare sulla porta seriale.

Il qualcosa in questione è il singolo carattere (byte) inviato dallo script di Python. Ricordiamo che esso contiene, in origine, un numero intero che, moltiplicato per 100, deve dare l'intervallo nominale in μs tra due istanti successivi di campionamento. Il *numero* di μs di questo intervallo è contenuto nella variabile `delays` che è costruita dalla lettura della porta seriale sfruttando un

di istruzione tanto misteriose quanto utili (le spiegazioni relative, non date qui, si trovano facilmente in rete): esse consentono ad Arduino di operare la digitalizzazione (quasi) al massimo del rete di campionamento possibile, che è dell'ordine di alcune decine di μs . Senza entrare troppo nei dettagli, lo scopo delle istruzioni è di istruire i registri interni al microcontroller in modo che esso sia in un certo senso “over-clocked” rispetto alle condizioni ordinarie. Naturalmente per l'esperienza presente questa specifica è del tutto inutile, visto che non serve a niente campionare “ad alta velocità” un segnale costante, però essa è mantenuta per analogia con quanto faremo nella maggior parte dei nostri impieghi sperimentali di Arduino.

La parte di sketch che riguarda l'inizializzazione è la seguente:

trucchettino. Infatti quello che originariamente, nelle nostre intenzioni, era un numero intero, è stato necessariamente convertito in un byte, ovvero in un carattere ASCII, prima di essere inviato attraverso porta seriale dal computer a Arduino. Per essere riconvertito in intero esso deve essere decodificato, operazione che viene effettuata con l'istruzione `Serial.read`. I numeri interi sono codificati ASCII in ordine nell'intervallo compreso tra il decimale 48, corrispondente al carattere '0', e il 57, corrispondente al carattere '9'. Di conseguenza la decodifica, per esempio, del carattere '5' dà luogo al numero intero 53. Dunque “sottrarre lo '0'”, che è codificato con il decimale 48, come fatto nello sketch, consente di ricavare il numero originario ($53 - 48 = 5$). Esso poi va, come detto, moltiplicato per 100 allo scopo di definire la variabile `delays`, che contiene il numero di μs che deve nominalmente intercorrere tra una digitalizzazione e la successiva.

Di seguito, dopo aver svuotato il buffer della porta seriale, la variabile `start` viene posta a 1 e l'acquisizione comincia. Per scopi puramente precauzionali, e probabilmente inutili, prima di iniziare le misure viene imposta

ad Arduino una pausa di 2000 ms attraverso l'istruzione `delay(2000)` (l'unità di misura è qui ms). Questa pausa tranquillizza nei confronti di possibili intasamenti nel funzionamento del microcontroller, in particolare nella fase di trasferimento dati via porta seriale USB (credo sia possibile ridurne la durata in caso di necessità).

A questo punto inizia il ciclo di misure. La digitalizzazione del segnale sulla porta di ingresso indicata dalla variabile `analogPin` avviene attraverso l'istruzione `analogRead(analogPin);`, che ritorna un intero compreso tra 0 e 1023. Notate che, prima di iniziare le misure "vere e proprie" (quelle che vengono effettivamente registrate), lo sketch prevede un ciclo di due misure a vuoto. Lo scopo è di minimizzare l'acquisizione di *artefatti*, cioè misure falsate, dovuti alla presenza segnali spuri generati all'interno di Arduino all'inizio delle misure. Il problema non è atteso avere effetti rilevanti per la presente esperienza, ma conviene prevedere le due misure "a vuoto" per uniformità con quanto faremo in futuro.

Quindi, subito prima di avviare il ciclo di misure vere e proprie (da registrare), Arduino misura il suo tempo interno (in unità di μs) e lo mette nella variabile `StartTime`, usando l'istruzione `StartTime=micros();`; il tempo interno scorre ciclicamente a partire dall'istante in cui il programma è stato lanciato e il valore che viene qui registrato verrà poi sottratto alle misure di tempo eseguite in corrispondenza delle digitalizzazioni, in modo da costruire un time stamp riferito sempre (nominalmente) all'inizio delle acquisizioni.

Finalmente ha inizio il ciclo di misure, che, in questo esempio, si svolge su 256 digitalizzazioni. L'istruzione che avvia il ciclo è `for(i=0;i<256;i++)`, con

```
// Istruzioni del programma
void loop()
{
    if (Serial.available() >0) // Controlla se il buffer seriale ha qualcosa
    {
        delays = (Serial.read()-'0')*100; // Legge il byte e lo interpreta come ritardo
        Serial.flush(); // Svuota la seriale
    start=1; // Pone il flag start a uno
    }
    if(!start) return // Se il flag e' start=0 non esegue le operazioni qui di seguito
                    // altrimenti le fa partire (quindi aspetta di ricevere l'istruzione
                    // di partenza
    delay(2000); // Aspetta 2000 ms per evitare casini
    for(i=0;i<2;i++) // Fa un ciclo di due letture a vuoto per "scaricare" l'analogPin
    {
        V[i]=analogRead(analogPin);
    }
    StartTime=micros(); // Misura il tempo iniziale con l'orologio interno
    for(i=0;i<256;i++) // Loop di misura
    {
        t[i]=micros()-StartTime; // Legge il timestamp e lo mette in array t
        V[i]=analogRead(analogPin); // Legge analogPin e lo mette in array V
        delayMicroseconds(delays); // Aspetta tot us
    }
}
```

ovvia sintassi; le istruzioni del ciclo sono comprese tra parentesi graffe. Il risultato delle misure va negli elementi i-esimi degli array `V[i]`, grazie all'istruzione `V[i] = analogRead(analogPin);`, e `t[i]`, grazie all'istruzione `t[i]=micros()-StartTime;`, che, come detto prima, determina il tempo a partire dall'istante iniziale immagazzinato in precedenza nella variabile `StartTime`. Queste istruzioni sono seguite dalla `delayMicroseconds(delays);`, che temporizza il campionamento su intervalli distanti temporalmente per il valore impostato [7]. Attraverso quanto descritto in seguito verificheremo la corrispondenza tra valore effettivo e nominale di tale intervallo.

Terminata l'acquisizione dei 256 punti sperimentali, ovvero la costruzione del record composto da 256 copie di dati (d.d.p. in digit e tempo in μs) comincia il ciclo di scrittura sulla porta seriale. Il modo con cui essi sono scritti non è molto elegante, ma garantisce di avere files in formato testo che possono facilmente essere letti da Python. I dati sono organizzati in righe contenenti, nell'ordine, il valore i-esimo dell'array `t[i]`, uno spazio, il valore i-esimo dell'array `V[i]`. Al termine di ciascuna riga si "va a capo" per iniziare una nuova; questo è realizzato dall'istruzione `Serial.println(V[i]);` (l'istruzione che serve per scrivere i dati senza andare a capo è `Serial.print`).

Alla fine di questo ciclo di scrittura la variabile `flag` viene annullata, in modo da uscire dall'acquisizione, e la porta seriale viene svuotata.

La corrispondente sezione di sketch è riportata nel seguente (l'intero sketch si trova in rete e nei computer di laboratorio sotto il nome `ardu2016.ino`):

```

for(i=0;i<256;i++) // Loop per la scrittura su porta seriale
{
    Serial.print(t[i]); // Scrive t[i]
    Serial.print(" ");
    Serial.println(V[i]); // Scrive V[i] e va a capo
}
start=0; // Annulla il flag
Serial.flush(); // Pulisce il buffer della porta seriale (si sa mai)
}

```

V. ANALISI DI DATI ESEMPIO

L'esperimento descritto consente, in poche parole, di ottenere un campione di misure di una grandezza, la d.d.p. presente tra pin A0 e GND, supposta costante, e un campione di misure degli istanti a cui la grandezza è stata digitalizzata. Anche se l'analisi non è particolarmente significativa dal punto di vista della fisica coinvolta, questi campioni possono essere trattati con metodi statistici (costruzione dell'istogramma delle occorrenze, calcolo di media e deviazione standard sperimentale). Lo scopo principale è quello di stimare l'incertezza da associare alle misure di d.d.p. digitalizzata e di tempo condotte da Arduino nelle condizioni tipiche dei nostri esperimenti.

Dunque qui di seguito vengono riportati esempi e commenti, rimandando alle sezioni successive per la discussione delle operazioni di calibrazione del digitalizzatore (che, invece, sono argomento principale dell'esperienza pratica di laboratorio).

A. Campione di misure digitalizzate

La Fig. 2 mostra un esempio di campione ottenuto registrando 8 blocchi consecutivi di 256 misure (usando nacqs=8 nello script di Python), per un totale di 2048 dati, in presenza di una d.d.p. $\Delta V = (2.65 \pm 0.02)$ V, misurata con il multimetro digitale. La lettura del multimetro rimaneva costante (entro l'errore) al distacco della porta di ingresso di Arduino, per cui per questa misura la resistenza di ingresso di Arduino è stata considerata sufficientemente alta da produrre effetti del tutto trascurabili, come atteso considerando la resistenza di ingresso del digitalizzatore, nominalmente pari a 100 Mohm.

Il pannello superiore riporta il valore della d.d.p. digitalizzata (dunque l'unità di misura è digit, secondo la nostra convenzione) in funzione del numero progressivo della misura. Ai dati sperimentali è associata un'*incertezza convenzionale* pari a ± 1 digit. Come si vede, la misura è stabile: l'intero campione è infatti costituito dal valore 540 digit, con deviazione standard sperimentale nulla.

Per completezza, il pannello inferiore mostra l'*istogramma delle occorrenze* dello stesso campione: si vede che un solo bin, quello corrispondente alla lettura 540 digit, è popolato. Per realizzare l'istogramma con Python esistono diverse possibilità: la più semplice con-

siste nell'uso dell'istruzione `pylab.hist`, che provvede a costruire e visualizzare l'array di istogramma delle occorrenze. Si ricorda che questa istruzione contiene un certo numero di argomenti: `pylab.hist(sample, bins = xx, range = (*,*), histtype = ??, color = ??, normed=??)`, dove `sample` è l'array che si intende analizzare e il resto è sufficientemente auto-esplicativo. Per avere una corretta rappresentazione occorre aggiustare il range e il numero di bin in modo tale che *i vari bin corrispondano a numeri interi*, che sono quelli effettivamente misurati: infatti è sicuro che bin "frazionari" hanno popolazione nulla. Inoltre, essendo interessati all'istogramma delle occorrenze, *non* è necessario, e anzi è sconsigliato, arrangiarsi per ottenere l'istogramma delle frequenze (normalizzate).

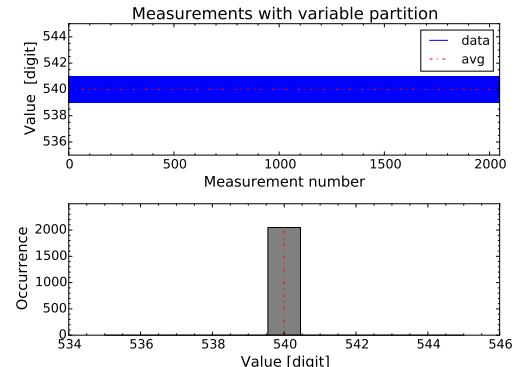


Figura 2. Esempio di analisi del campione di dati acquisito per $\Delta V = (2.65 \pm 0.02)$ V, misurata con il multimetro digitale. Il pannello superiore riporta il grafico delle misure, a cui è stata attribuita un'incertezza convenzionale ± 1 digit, il pannello inferiore riporta l'istogramma delle occorrenze. Il valore medio digitalizzato è 540 digit e la deviazione standard sperimentale è nulla, secondo quanto descritto nel testo.

La situazione considerata, in cui si registra una deviazione standard sperimentale nulla, è piuttosto comune. In linea di principio, in un esperimento come quello qui condotto ci si aspetta che le misure abbiano una qualche distribuzione, che presumibilmente approssima la distribuzione normale (Gaussiana) all'aumentare delle dimensioni del campione. Qui la distribuzione non può essere ricostruita e questo si verifica semplicemente perché la sensibilità della misura (il valore fisico che corrisponde

al singolo digit) non è sufficiente. Evidentemente, non è possibile stimare l'incertezza dalla “misura” (della media) dalla deviazione standard sperimentale, che ha un valore inferiore alla massima sensibilità. Di conseguenza sceglieremo di individuare un'incertezza arbitraria, che convenzionalmente poniamo pari a ± 1 digit. Notate che, così facendo, possiamo ancora sostenerne come tale incertezza abbia un'origine prevalentemente statistica, ma sicuramente ne stiamo compiendo una sovrastima. D'altra parte, come avremo modo di verificare in ulteriori esperienze, il digitalizzatore di Arduino può anche avere incertezze di origine sistematica, specialmente quando impiegato per acquisire d.d.p. variabili nel tempo.

Ricordate poi che la lettura automatizzata di dati non esercita (non può esercitare, a meno di introdurre opportune strategie) alcuna forma di controllo sui valori effettivamente letti. È possibile che alcune delle letture eseguite da Arduino siano falsate da artefatti di varia natura, per esempio da “disturbi” sull'alimentazione o nello stadio di ingresso del digitalizzatore [8]. Queste misure false contribuiscono all'istogramma, pur essendo non necessariamente rappresentative della distribuzione Gaussiana che ci aspetteremmo (un esempio è mostrato nella sezione seguente). È interessante notare che la frequenza delle misure false può dipendere dalle condizioni di funzionamento di Arduino: negli esempi mostrati in questa nota, Arduino era collegato a un computer portatile alimentato a batteria. L'uso dei computer di laboratorio, alimentati dalla rete (e collegati alla linea di terra) può sicuramente aumentare la probabilità di registrare artefatti.

1. Operazione con riferimento interno a 1.1 V

Normalmente Arduino lavora usando una tensione di riferimento $V_{max} \sim 5$ V generata a partire da quella di alimentazione (l'alimentazione nel nostro caso è fornita dalla presa USB del computer), che in sostanza rappresenta la portata della misura di d.d.p.: su questo argomento torneremo in seguito, quando affronteremo la calibrazione del digitalizzatore. Tuttavia è ovvio che V_{max} determina la sensibilità dello strumento.

Esiste un'opzione, che si richiama mettendo nel blocco di inizializzazione dello sketch l'istruzione `analogReference(INTERNAL);`, che ordina ad Arduino di impiegare come riferimento una tensione generata internamente, di valore nominale $V_{max} = 1.1$ V (l'incertezza non è nota). Evidentemente l'uso di questa istruzione modifica la sensibilità del digitalizzatore, che diventa di circa 1 mV, 5 volte minore di quella “ordinaria”. Questa possibilità apre la strada per ottenere un campione con qualche distribuzione osservabile.

L'opzione è attivata nello sketch `ardu_1V1_2016.ino` disponibile in rete e nei computer di laboratorio. Ovviamenete se si usa questo sketch occorre *tassativamente* che la d.d.p. in ingresso alla porta analogica A0 sia $\Delta V \leq 1.1$ V: ciò si ottiene, per esempio, inserendo $R_1 = 6.8$ kohm (nominali) nel circuito di Fig. 1.

La Fig. 3 mostra un esempio di campione ottenuto con questo sketch, usando $\Delta V = (964 \pm 5)$ mV: si vede come questa volta la sensibilità della misura sia sufficiente per apprezzare delle variazioni, anche se la distribuzione suggerita dall'istogramma, fortemente asimmetrico, non assomiglia affatto a una Gaussiana. Per l'esempio considerato si ottiene un valore medio dei conteggi di 904.6 digit e una deviazione standard sperimentale di 7.6 digit. La “strana” distribuzione registrata potrebbe essere dovuta alla presenza di artefatti, secondo quanto accennato in precedenza.

Come ultima osservazione molto rilevante (e altrettanto ovvia), notiamo che in questi esperimenti la d.d.p. da misurare è ritenuta costante, ma non si ha nessuna garanzia che essa lo sia. In altre parole, le fluttuazioni registrate nel campione potrebbero avere luogo anche nel circuito (alimentatore e partitore) usato per generare la d.d.p. stessa, come è molto probabile che si verifichi per esempio a causa delle scarse stabilità meccanica, e quindi elettrica, del potenziometro.

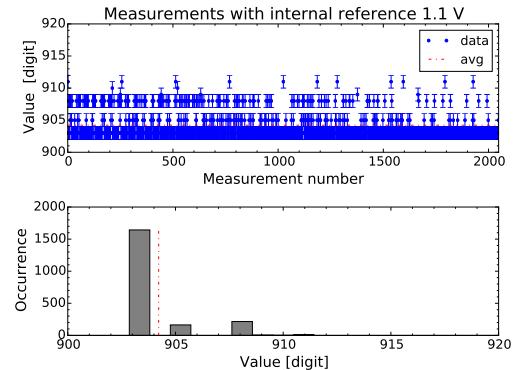


Figura 3. Analogo di Fig. 2 ottenuto utilizzando la tensione di riferimento interna $V_{ref} = 1.1$ V (nominali) e misurando con il multimetro digitale $\Delta V = (964 \pm 5)$ mV. Il valore medio digitalizzato è 904.6 digit e la deviazione standard sperimentale è 7.6 digit.

B. Campione degli intervalli di campionamento

Le acquisizioni di questa esperienza si prestano anche a un altro tipo di analisi. Infatti possiamo esaminare la misura dei tempi come effettuata da Arduino (registrata in unità di μs nella prima colonna del file ottenuto) allo scopo di determinare l'intervalllo effettivo di campionamento Δt , cioè l'intervalllo temporale tra due misure successive. Lo scopo è quello di confrontare il risultato con il valore nominale Δt_{nom} impostato nello script di Python (in questo esempio, $\Delta t_{nom} = 500 \mu\text{s}$) e di dedurre la deviazione standard sperimentale del campione degli intervalli di campionamento.

Arduino ha un clock interno primario realizzato con un oscillatore al quarzo di frequenza 16.000 MHz. Dunque virtualmente la massima accuratezza nella misura dei

tempi dovrebbe essere di circa 60 ns. Tuttavia le operazioni di digitalizzazione richiedono numerosi cicli di clock primario, il cui numero, oltre tutto, può essere influenzato da fluttuazioni e latenze dovute alla contemporaneità di altre operazioni compiute dal microcontroller assieme alla digitalizzazione (per esempio tutte quelle operazioni necessarie a gestire la comunicazione con la memoria SRAM). Dalle specifiche di Arduino si sa che i tempi sono determinati, e quindi anche misurati, con una risoluzione di 4 μ s, decisamente peggiore di quella virtualmente consentita dalla frequenza del clock primario. Dunque appare ragionevole impiegare in prima battuta un'incertezza convenzionale $\delta t = 4 \mu$ s per la misura dei tempi. L'analisi del campione ci consentirà di verificare la validità di questa scelta, secondo quanto discusso nel seguito.

La Fig. 4 mostra nel pannello superiore gli intervalli di tempo misurati Δt in funzione del numero progressivo di misura, corredati dell'incertezza appena specificata; il pannello inferiore rappresenta l'istogramma delle occorrenze per il campione considerato. Notate che produrre il campione dal file è un'operazione non del tutto banale: infatti, detti t_j i tempi misurati da Arduino, da ogni blocco di 256 misure possono essere estratti (256 – 1) valori $\Delta t_j = t_{j+1} - t_j$, e la produzione di un unico array contenente i dati per l'intera acquisizione richiede attenzione nello scrivere un corretto algoritmo (questa parte può sicuramente essere considerata facoltativa nell'esperienza pratica).

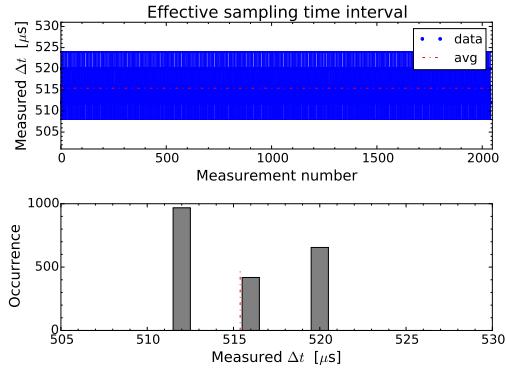


Figura 4. Esempio di analisi del campione di intervalli di campionamento Δt costruito come discusso nel testo. Il pannello superiore riporta il grafico delle misure, a cui è stata attribuita un'incertezza convenzionale $\pm 4 \mu$ s, il pannello inferiore riporta l'istogramma delle occorrenze. Il valore medio digitalizzato è 515.4 μ s e la deviazione standard sperimentale è 3.5 μ s.

Come si osserva in figura, l'intervallo Δt è diverso dall'impostazione nominale $\Delta t_{nom} = 500 \mu$ s: il valore medio ottenuto è infatti 515.4 μ s. La discrepanza può facilmente essere attribuita ai tempi di latenza del microcontroller e, soprattutto, al tempo minimo effettivo necessario affinché la digitalizzazione sia conclusa, altrove definito Δt_{dig} , che dalle misure risulta di circa 15 μ s. La deviazione standard sperimentale ottenuta dal campione è 3.5

μ s: questo valore può essere considerato come un'indicazione dell'incertezza associata alla misura del tempo da parte di Arduino nelle condizioni del nostro esperimento. Notate tuttavia che la distribuzione mostrata nell'istogramma è anche in questo caso tutt'altro che Gaussiana. Il risultato suggerisce comunque che l'incertezza dedotta dalle specifiche (4 μ s) è simile (leggermente superiore, nel nostro esempio) a quella determinata esaminando il campione.

VI. CALIBRAZIONE DI ARDUINO

Nelle nostre esercitazioni pratiche Arduino viene impiegato principalmente come misuratore di d.d.p.: anche se in qualche occasione sarà sufficiente per i nostri scopi conoscere le tensioni misurate in unità arbitrarie di digitalizzazione (digit), spesso sarà necessario convertire tali unità in unità fisiche (V). Per questo è necessario eseguire una *calibrazione*.

In termini generali, è evidente che la calibrazione dipende dalle caratteristiche della specifica scheda Arduino impiegata e dal valore di V_{max} che, in condizioni di operazione ordinarie (senza cioè scegliere il riferimento interno da 1.1 V nominali), dipende dall'alimentazione che Arduino riceve tramite USB. Dunque in linea di principio la calibrazione andrebbe ripetuta a ogni impiego di Arduino. Fortunatamente, come specificheremo nella prossima sezione, esiste un modo approssimato e molto rapido (*calibrazione alternativa*) per determinare un fattore di calibrazione, cioè di conversione tra digit e V. Qui, però, intendiamo operare come il faut, facendo finta di essere i costruttori di uno strumento di misura e di trovarci impegnati nella sua calibrazione.

La calibrazione si fa, normalmente, *per confronto*, cioè attraverso lettura di un campione di unità di misura calibrato. Noi non disponiamo di un campione di tensione calibrato e il meglio che possiamo fare è produrre una d.d.p. (con generatore e partitore di tensione, come visto in precedenza), misurarla con il multimetro digitale, la cui calibrazione è nota, benché affetta da una incertezza sicuramente non trascurabile, e quindi usare la lettura del multimetro con la sua incertezza come campione di riferimento. Basare la calibrazione su una singola misura implica, inevitabilmente, di assumere una *proporzionalità diretta* tra digit e V, cioè un comportamento perfettamente lineare del digitalizzatore. Spulciando nelle specifiche del microcontroller si vede come questa affermazione possa essere non completamente corretta: infatti le specifiche suggeriscono come possa esserci un *offset* diverso da zero nella digitalizzazione. In altre parole, a una tensione di ingresso nulla entro le incertezze, $\Delta V = 0$, potrebbe corrispondere una lettura in digit diversa da zero, o viceversa. Inoltre la stessa dipendenza lineare tra digit e V deve, in linea di principio, essere verificata sperimentalmente entro le incertezze della misura.

Per rispondere a queste esigenze occorre eseguire un esperimento in cui vengono raccolte le misure digitaliz-

zate, qui indicate con il simbolo X , in corrispondenza di diversi valori della d.d.p. ΔV in ingresso. I dati possono quindi essere interpretati tramite un best-fit secondo la seguente funzione modello lineare, suggerita dalle specifiche di Arduino:

$$\Delta V = \alpha + \beta X, \quad (1)$$

dove i parametri α e β possono essere determinati tramite best-fit.

Nell'esempio da me realizzato ho usato lo script `ardu2016.py` e lo sketch `ardu2016.ino` per costruire campioni di 256 misure (`nacqs = 1` nello script) acquisiti in corrispondenza di diversi valori ΔV_j realizzati ruotando in j -diverse posizioni l'alberino del potenziometro. Dalle indicazioni che lo script produce sulla console ho determinato il valore medio X_j della digitalizzazione effettuata. Come incertezze, per la misura con il multimetro ho seguito le consuete indicazioni del manuale, per i dati digitalizzati ho usato l'incertezza convenzionale ± 1 digit, a meno che la deviazione standard sperimentale del campione, indicata sempre sulla console, non fosse maggiore.

Il risultato delle misure di calibrazione è riportato in Fig. 5 assieme alla retta ottenuta dal best-fit, i cui risultati sono

$$\alpha = (19 \pm 3) \text{ mV} \quad (2)$$

$$\beta = (4.87 \pm 0.02) \text{ mV/digit} \quad (3)$$

$$\chi^2/\text{ndof} = 43/17 \quad (4)$$

$$\text{norm.cov.} = -0.65 \quad (5)$$

$$\text{absolute_sigma} = \text{False}; \quad (6)$$

notate che nel best-fit si è considerata l'incertezza sui valori digitalizzati X_j attraverso la consueta procedura (errore equivalente determinato con propagazione e somma in quadratura). Come indicato dal grafico dei residui normalizzati, l'accordo tra dati e modello è scarso soprattutto per bassi valori di ΔV , dove probabilmente la risposta del digitalizzatore tende ad essere non lineare (siete invitati a verificare di quanto l'accordo possa migliorare usando, per esempio, una dipendenza polinomiale di ordine superiore).

A. Calibrazione alternativa

La procedura di calibrazione alternativa (così chiamata per nostra convenzione) sfrutta una relazione di proporzionalità diretta tra d.d.p. in V e valore digitalizzato X , secondo la

$$\Delta V = \xi X. \quad (7)$$

Dunque in questa relazione la presenza dell'offset (rappresentato dal parametro α sopra determinato) è trascurata.

Evidentemente, se fosse nota la tensione di riferimento V_{max} , che corrisponde al massimo valore di d.d.p. che può

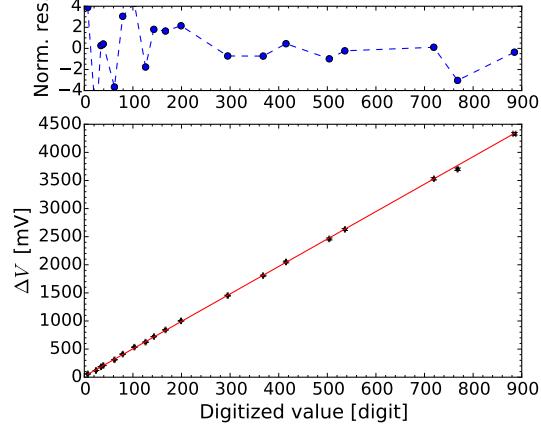


Figura 5. Risultato delle misure di calibrazione discusse nel testo: il pannello inferiore mostra dati e retta ottenuta dal best-fit, quello superiore il grafico dei residui normalizzati.

essere digitalizzata e dunque a $X = 1023$ digit, sarebbe possibile ricavare ξ dalla semplice relazione

$$\xi = \frac{V_{ref}}{1023}. \quad (8)$$

Il valore di V_{max} è però noto (in forma nominale) solo usando il riferimento interno da 1.1 V. Negli altri casi, si può utilizzare un'utile scorciatoia suggerita dal tanto materiale disponibile in rete. Infatti V_{max} è attesa corrispondere anche alla massima tensione che può essere prodotta da Arduino sulle sue porte digitali di uscita. Quindi misurando con il multimetro digitale questa tensione è possibile conoscere in maniera immediata ξ e ottenere una calibrazione (approssimativa, ma spesso ragionevole per i nostri scopi) dello strumento.

Nello sketch qui utilizzato, come abbiamo già sottolineato, Arduino viene istruito a usare la porta corrispondente al pin 7 (collegato con filo arancione/rosso a una boccola volante rossa) come uscita digitale e a porla a "livello alto", cioè a tenerla "accesa" per tutta la durata dell'esperienza. La nostra ipotesi, supportata da molte informazioni disponibili in rete, è che la tensione misurata fra questa porta e la linea di terra in queste condizioni, ΔV_{pin7} , sia pari nominalmente a V_{max} . Nell'esempio qui riportato, si è ottenuto $\Delta V_{pin7} = (4.94 \pm 0.03)$ V, a cui corrisponde $\xi = (4.83 \pm 0.03)$ mV/digit: questo valore è compatibile con quello del parametro β prima determinato con il best-fit, circostanza che fornisce una prima indicazione qualitativa sulla validità della procedura di calibrazione alternativa.

Una valutazione un po' più quantitativa può essere ottenuta come descritto qui nel seguito. Dal best-fit lineare della calibrazione è possibile determinare il valore previsto ΔV_{prev} corrispondente a una arbitraria lettura digitalizzata X (ΔV_{prev} è quindi da intendersi come una funzione di X). In questa previsione è utile tenere conto della covarianza dei due parametri di fit, secondo quanto

già conosciamo; ricordiamo infatti che vale la relazione

$$\delta V_{prev} = \sqrt{C_{11}^2 + C_{22}X^2 + 2C_{12}X}, \quad (9)$$

dove δV_{prev} è l'incertezza della previsione, C_{ij} sono gli elementi della *matrice di covarianza* ottenuti dal best-fit lineare (numerico), X è il valore della lettura digitalizzata. La stessa operazione possiamo farla per la previsione basata sulla calibrazione alternativa, $\Delta V_{prev,alt} = \xi X$: stavolta l'impiego di un unico parametro rende inutile preoccuparsi della covarianza (che non è definita in questo caso), per cui, con ovvio significato dei simboli, possiamo porre $\delta V_{prev,alt} = \Delta \xi X$.

A questo punto possiamo costruire le cosiddette “curve di confidenza” (nome convenzionale), cioè le funzioni $\Delta V_{prev} \pm \delta V_{prev}$ e $\Delta V_{prev,alt} \pm \delta V_{prev,alt}$; il risultato è mostrato in Fig. 6, dove si è impiegata la rappresentazione logaritmica per mettere meglio in evidenza le piccole differenze in valore assoluto (per questo grafico, X è stata costruita con un array equispaziato logaritmicamente). Se i due metodi di calibrazione portassero a risultati in accordo fra loro entro le incertezze, le “bande di confidenza” dovrebbero essere una dentro l'altra. Apparentemente (cioè con la risoluzione consentita dal grafico) questo non si verifica sempre e, in particolare, per bassi valori di ΔV , ovvero di X , ci sono discrepanze facilmente apprezzabili. Infatti è ovvio che, pur se l'intercetta del modello lineare (parametro α del fit) è piccola in valore assoluto, essa produce degli effetti nella calibrazione di valori digitalizzati corrispondenti a pochi digit.

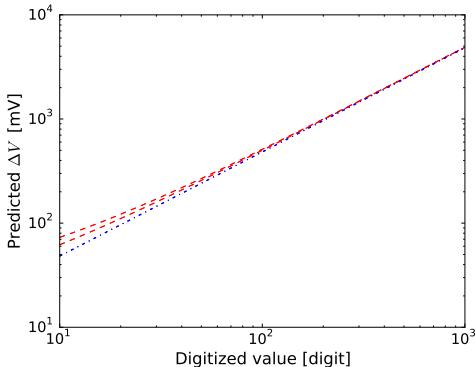


Figura 6. “Curve di confidenza” per i valori ΔV previsti in funzione della lettura digitalizzata X : secondo quanto discusso nel testo, le linee tratteggiate rosse rappresentano il risultato ottenuto considerando la calibrazione con best-fit lineare, quelle punto-linea blu la calibrazione alternativa.

Allo scopo di ottenere un'indicazione quantitativa, è utile creare la funzione ∂ , simbolo con cui indichiamo la *massima* discrepanza (in valore assoluto) tra i risultati delle due calibrazioni, normalizzata rispetto alla lettura ΔV . Dal punto di vista matematico, si può porre, per esempio,

$$\partial = \frac{\max\{|\Delta V_{prev} - \Delta V_{prev,alt}|\}}{\Delta V_{prev}}. \quad (10)$$

Il grafico di questa grandezza è rappresentato in Fig. 7: si vede come la discrepanza relativa ∂ tra i due metodi di calibrazione sia tutt'altro che trascurabile per bassi valori digitalizzati, ma anche come essa scenda sotto il 5% quando la lettura del digitalizzatore è superiore al centinaio.

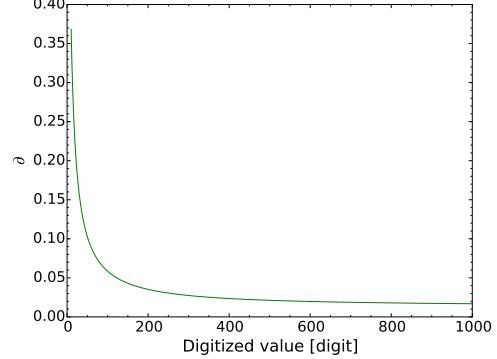


Figura 7. Discrepanza relativa ∂ tra i due metodi di calibrazione, come definita nel testo, in funzione del valore digitalizzato.

VII. QUALCHE APPARENTE STRANEZZA

Da ultimo, in questa sezione si riporta un'osservazione che può essere facilmente registrata in modo involontario, ma che qui è costruita apposta e che anche voi siete invitati a registrare volontariamente. Può succedere talvolta che si esegua una digitalizzazione mantenendo scollegato l'ingresso di Arduino (ovvero collegato su un carico resistivo elevato: si ottengono risultati simili a quelli qui mostrati anche chiudendo l'ingresso di Arduino, il pin A0, su una resistenza dell'ordine del Mohm). In queste condizioni ci si potrebbe aspettare di ottenere una lettura costantemente nulla, a parte piccole fluttuazioni.

Un risultato esempio è mostrato in Fig. 8 (la figura è costruita come Fig. 2 e quindi riporta anche l'istogramma delle occorrenze): si vede che le letture sono ben diverse da zero e che la loro distribuzione popola diversi bins e non solo quello corrispondente alla lettura nulla.

L'interpretazione è piuttosto immediata: quelle registrate sono letture dovute a “disturbi” che vengono raccolti dall'ingresso del digitalizzatore. Poiché le digitalizzazioni sono pressoché equispaziate nel tempo, il grafico suggerisce un andamento periodico, o quasi-periodico, dei disturbi e un'analisi che tiene conto anche del tempo di digitalizzazione indica in circa 50 Hz la frequenza (prevalente) dei disturbi. Allora è evidente che essi hanno origine nella corrente di rete, che, essendo alternata, dà luogo, attraverso meccanismi che vi saranno chiari proseguendo nel corso, a una d.d.p. anche alternata e alla stessa frequenza di 50 Hz (o un suo multiplo). In futuro indicheremo talvolta fenomeni di accoppiamento dei disturbi agli strumenti di misura come *rumori di pick-up* e

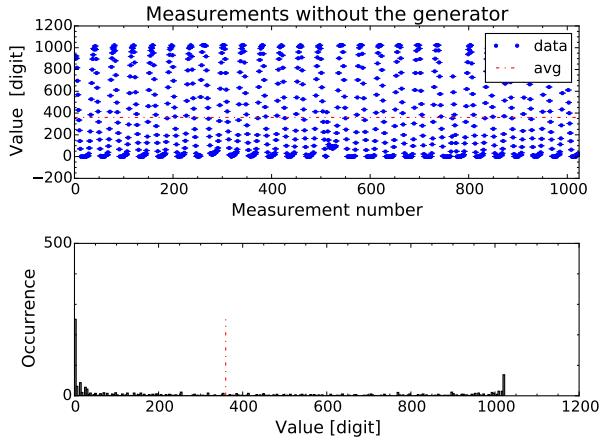


Figura 8. Analogo di Fig. 2 dove, però, il generatore è stato scollegato, lasciando “volante” la boccola collegata al pin A0 di Arduino, ovvero *flottante* la d.d.p. in ingresso al digitalizzatore.

potrebbe essere interessante, fra qualche mese, esaminare il record così costruito con il metodo della *trasformata di Fourier numerica*.

Della presenza di queste fluttuazioni non ci si rende

generalmente conto usando altri strumenti di misura: lasciando scollegato il multimetro digitale configurato come voltmetro (in corrente continua), non si osservano variazioni della lettura che possano essere ascritte a questo tipo di disturbi. Ci sono diversi motivi che possono spiegare lo specifico comportamento di Arduino confrontato con quello del voltmetro digitale: quest’ultimo, infatti, ha un tempo di refresh del display piuttosto lungo (sicuramente più lungo degli intervalli di campionamento qui impiegati) e, in pratica, esso media a zero, ovvero “filtra”, i disturbi alternati. In secondo luogo, il layout del multimetro digitale è certamente più accurato di quello della scheda Arduino, almeno come la usiamo noi, per cui i disturbi alternati potrebbero essere “schermati” e attenuati in ampiezza. Inoltre il multimetro è alimentato a batteria e non ha di per sé alcun collegamento fisico con la rete elettrica, o la linea di terra. Infine, la resistenza di ingresso di Arduino è superiore rispetto a quella (già molto alta) del multimetro digitale usato come voltmetro, e questo potrebbe aumentare l’ampiezza della d.d.p. creata dal disturbo.

In ogni caso, ricordate sempre che non collegare nulla a uno strumento di misura non significa necessariamente porre pari a zero la grandezza in ingresso: azzerare il segnale letto implica infatti collegare l’ingresso alla linea di massa, o di terra.

- [1] In seguito a confusi sviluppi della situazione societaria di Arduino, esiste anche altre denominazioni per la scheda, tra cui “Genuino”.
- [2] La situazione sarà diversa in esperienze future, nelle quali sarà indispensabile eseguire misure *sincrone*: allo scopo, verranno implementate strategie in grado di triggerare (brutto termine che vi diventerà molto familiare) l’acquisizione in contemporanea con il verificarsi di determinati eventi.
- [3] Il motivo è banale. Il microcontroller di Arduino, come qualsiasi CPU, ha ritardi, o tempi di latenza, durante l’esecuzione di un ciclo di programma che non possono essere determinati a priori. Infatti, oltre ad eseguire il programma, la CPU deve preoccuparsi di gestire il suo stesso funzionamento, cioè di controllare, per esempio, lo stato della memoria, la sua alimentazione, la presenza di segnali sulle porte, l’accensione o lo spegnimento di segnali interni (“interrupts”), etc.. Di conseguenza il controllo sul timing delle varie operazioni, inclusi i cicli di ritardo inseriti nel programma, è affetto da incertezza interpretabile come statistica.
- [4] Per favore, tenete conto che il potenziometro è, per sua natura, un dispositivo delicato e poco affidabile. Infatti il contatto strisciante può facilmente funzionare in modo non corretto, dando luogo a una resistenza diversa da quella attesa per una data posizione dell’alberino. Inoltre in certe condizioni è sufficiente sfiorare la manopola fissata sull’alberino per ottenere variazioni poco controllate della resistenza.
- [5] A parte gli ovvi motivi di disponibilità di tempo e di capa-

cità di elaborazione dati, ci sono buone ragioni per evitare acquisizioni troppo lunghe, cioè ripetute su più di *qualche* ciclo. Esse risiedono principalmente nel fatto che il sistema qui impiegato, come la maggior parte dei sistemi fisici, soffre di variazioni delle condizioni di funzionamento (*drifts*) a medio termine, per esempio sulla scala dei minuti. Queste variazioni possono essere legate a diverse cause fisiche: di norma, per i nostri esperimenti la principale è la variazione di temperatura, che si sviluppa tipicamente su queste scale temporali.

- [6] Nell’elettronica digitale, gli stati, o livelli, alti o bassi corrispondono rispettivamente agli 0 e 1 della logica binaria. I valori di d.d.p. corrispondenti possono essere definiti secondo numerosi standard, il più comune dei quali è lo standard TTL (Transistor-Transistor Logic), che è anche impiegato per le porte digitali di Arduino. Esso fa corrispondere nominalmente il livello basso a una tensione compresa tra 0 e 0.4 V, e il livello alto a una tensione superiore a 2.4 V. L’implementazione di molti dispositivi, Arduino compreso, prevede che il livello basso corrisponda a una d.d.p. circa nulla, e quello alto a una d.d.p. di circa 5 V.
- [7] Visto come è realizzato lo sketch, è evidente il vantaggio di avere campionamenti nominalmente equispaziati nel tempo.
- [8] In futuro torneremo su questi “disturbi”, indicandoli talvolta come *rumore* prodotto dall’interazione con l’ambiente. Molto grossolanamente, infatti, chiameremo rumore tutte le fluttuazioni dei valori letti diverse da cause fisiche che possiamo controllare.

Circuiti complicati

francesco.fuso@unipi.it; <http://www.df.unipi.it/~fuso/dida>

(Dated: version 4 - FF, 13 ottobre 2016)

Questa nota riporta dei facili esempi di applicazione di alcune “regole” per la “soluzione” di circuiti elettrici. Tali regole rappresentano in sostanza la formalizzazione di considerazioni pressoché ovvie nell’elettrostatica e elettricità (definizioni, “principi” di conservazione, etc.), per cui non aggiungono conoscenza a quanto già noto. Tuttavia esse possono essere utili per determinare correnti e d.d.p. in circuiti complicati, specialmente quando sono presenti più generatori (di d.d.p. o di corrente), come spesso succede in elettronica.

I. NOMENCLATURA E REGOLE

Un circuito può essere genericamente indicato come la composizione di diversi componenti, uniti tra loro a formare uno o più sottocircuiti collegati e chiusi su se stessi attraverso fili (perfettamente) conduttori. I componenti possono in generale avere due o più elettrodi (o reofori, o fili di collegamento): per il momento conosciamo solo componenti a due elettrodi, detti anche *bipoli*, in particolare generatori di d.d.p., in questo contesto considerati ideali, e resistori. Pertanto gli esempi di questa nota saranno costruiti solo collegando resistori e generatori di d.d.p. tra loro.

Diamo un po’ di nomenclatura:

1. si chiama *nodo* il punto in cui si diramano le correnti, cioè la congiunzione tra tre o più fili;
2. si chiama *ramo* il tratto di circuito che congiunge due nodi (adiacenti);
3. si chiama *maglia* la successione di rami che si richiudono su se stessi.

In presenza di un circuito complicato, il primo obiettivo è quello di scomporlo in diverse maglie, ovvero diversi sotto-circuiti. Esiste un teorema che stabilisce che il numero delle maglie (indipendenti) N_m dipende da quello dei rami e dei nodi (rispettivamente N_r e N_n) secondo la relazione

$$N_m = N_r - N_n + 1 . \quad (1)$$

“Risolvere” un circuito equivale nella pratica a scrivere un sistema di equazioni algebriche (almeno per il momento, lineari) che contiene tante equazioni quante maglie indipendenti. Ognuna di queste equazioni lega tra loro grandezze elettriche di interesse, che per noi sono differenze di potenziale ΔV e intensità di corrente I .

Le equazioni devono essere scritte tenendo conto del modello, o *relazione costitutiva*, che descrive il funzionamento dei vari componenti, cioè, per noi al momento, la legge di Ohm [1]. Esistono poi due regoline, che qualche volta prendono il pomposo nome di “leggi di Kirchoff”:

$$\sum_i \Delta V_i = 0 \text{ su una maglia} \quad (2)$$

$$\sum_i I_i = 0 \text{ su un nodo ,} \quad (3)$$

dove le somme si intendono estese su tutte le d.d.p. ai capi degli elementi della maglia considerata, o su tutte le correnti che interessano il nodo considerato. Si intende che, nella scrittura data, devono essere impiegate opportune convenzioni per stabilire i segni di ΔV_i (rispetto a un verso convenzionalmente positivo di percorrenza della maglia) e di I_i (generalmente positivo/negativo a seconda che la corrente entra/esca dal nodo). Di questo ci renderemo facilmente conto svolgendo gli esercizi di esempio. Le Eqq. 2,3 hanno un’ovvia origine fisica legata rispettivamente al carattere additivo dei potenziali scalari statici (ricordatene la definizione attraverso integrale di linea) e alla conservazione della carica che per unità di tempo passa attraverso un nodo (il nodo non è né un pozzo, né una sorgente di carica).

Il modo con cui si scrivono le tante equazioni necessarie alla soluzione del circuito non è univoco. Qui useremo un metodo, generalmente molto efficiente, che si basa sull’Eq. 2 e che pertanto mescola in ogni singola equazione le correnti delle diverse maglie. Ci sono anche dei metodi che si basano sull’Eq. 3 e tanti metodi “misti”, che usano le due regole scritte sopra. Questi metodi misti sono quelli che in pratica si usano, in maniera più o meno immediata, quando si risolvono circuiti molto semplici, in cui è facile distinguere collegamenti in serie e parallelo.

Esiste poi un altro metodo molto utile nel caso in cui in un circuito siano presenti più generatori di d.d.p., o di corrente, che ugualmente serve per scrivere tante equazioni quante sono le maglie indipendenti del circuito. Esso si basa sul *principio di sovrapposizione*, valido grazie al carattere “lineare” della legge di Ohm, che viene applicato a correnti e d.d.p.. Le regoline su cui si basa questo metodo sono:

1. la soluzione del circuito può essere compiuta a passi successivi, cioè scrivendo tante equazioni in cui tutti i generatori *tranne uno* sono sostituiti da un *cortocircuito* (se generatore di d.d.p.) o circuito aperto (se generatore di corrente, caso che non esamineremo);
2. la soluzione “finale” si ottiene allora sovrapponendo, cioè sommando algebricamente, le d.d.p. e le correnti determinate nei vari passi.

Anche di queste regoline, almeno in parte, faremo uso in qualche esempio, in modo da chiarirne la ricetta pratica

di impiego. Per il momento osservate che sostituire un generatore di d.d.p. (ideale, in questo contesto) con un cortocircuito è in accordo con le specifiche dei generatori secondo il modello di Thévenin.

II. ESEMPIO MOLTO SEMPLICE

La Fig. 1 mostra un semplicissimo circuito costituito dal generatore di d.d.p. V_0 (ideale) e da tre resistenze R_{1-3} . La sua semplicità, in particolare il fatto che è presente un solo generatore, fa sì che a nessuna persona dotata di minimo buon senso venga in mente di impiegare le regole di cui alla sezione precedente. Infatti è ben più semplice risolvere il circuito basandosi su considerazioni immediate, o intuitive, le quali fanno in pratica uso di una mescolanza di tutte le varie regoline prima enunciate. È evidente che il circuito è un partitore di tensione realizzato dalla serie di R_1 con il partitore di corrente costituito dal parallelo di R_2 e R_3 . Di conseguenza è facile dedurre tutte le informazioni di interesse. Per esempio, la corrente I erogata dal generatore è $I = V_0/R_{tot}$, con $R_{tot} = R_1 + R_{2//}R_3 = (R_1R_2 + R_1R_3 + R_2R_3)/(R_2 + R_3)$. Se fossimo, ancora per esempio, interessati a conoscere la differenza di potenziale ΔV_2 ai capi di R_2 (naturalmente uguale a ΔV_3 ai capi di R_3), potremmo immediatamente affermare che essa è $\Delta V_2 = (R_{2//}R_3)I = V_0(R_2R_3)/(R_1R_2 + R_1R_3 + R_2R_3)$, e così via.

Tuttavia, al solo scopo di esercitarcisi, proviamo a ragionare nei termini prima stabiliti. Cominciamo individuando i nodi, che sono in numero $N_n = 2$, e i rami, che sono in numero $N_r = 3$ (nodi e rami sono marcati in figura con le lettere n_i e r_i). Dunque si hanno $N_m = 2$ maglie indipendenti. La scelta delle maglie non è univoca: decidiamo di usare le maglie indicate con $m_{1,2}$ in figura e stabilire che il loro verso di percorrenza è positivo quando le correnti fluiscono in senso orario. Questa scelta è ovvia considerando la disposizione dei poli del generatore (la corrente va dal positivo al negativo), che permette di stabilire il verso di percorrenza di m_1 , da cui anche quello di m_2 , che deve essere coerente.

Prima di procedere con la soluzione, conviene individuare quali componenti, o rami, si trovano *a comune* tra diverse maglie, cosa che ovviamente dipende dalla scelta delle maglie stesse. Per la scelta fatta, R_2 si trova ad essere interessata dalle correnti di tutte e due le maglie (con versi opposti). Di conseguenza la d.d.p. ai suoi capi dipenderà da *entrambi* le intensità di corrente delle due maglie [2].

Scriviamo ora le equazioni per i componenti delle due maglie seguendo le regole di Eq. 2 e partendo dal punto di circuito collocato “in basso a sinistra” in figura (questo giustifica i segni):

$$0 = V_0 - R_1 I_{m1} - R_2 I_{m1} + R_2 I_{m2} \text{ per la maglia } m_1 \quad (4)$$

$$0 = R_2 I_{m1} - R_2 I_{m2} - R_3 I_{m2} \text{ per la maglia } m_2, \quad (5)$$

dove abbiamo indicato con I_{m1} e I_{m2} le intensità di corrente delle due maglie. Osservate che i segni dei contri-

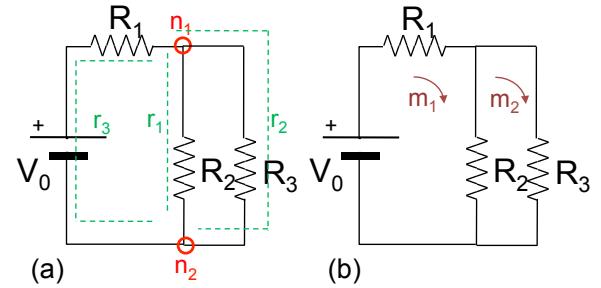


Figura 1. Schema del circuito considerato nel testo con indicati i nodi e i rami (a), e le maglie indipendenti utilizzate nella soluzione (b).

buti alla d.d.p. ai capi di R_2 devono essere opposti tra loro a causa della scelta dei versi di circolazione.

Personalmente trovo più conveniente scrivere le equazioni delle maglie in modo che l’eventuale presenza del generatore sia chiara guardando il primo membro delle equazioni stesse. In questo modo le Eqq. 4, 5 diventano, con qualche riaggiustamento di segno,

$$V_0 = R_1 I_{m1} + R_2 I_{m1} - R_2 I_{m2} \text{ per la maglia } m_1 \quad (6)$$

$$0 = R_2 I_{m1} - R_2 I_{m2} - R_3 I_{m2} \text{ per la maglia } m_2. \quad (7)$$

Supponendo di conoscere i valori di V_0 e delle resistenze, la soluzione del sistema delle due equazioni lineari algebriche consente di dedurre tutto quello che si vuole sul circuito in esame. Le intensità di corrente di maglia sono determinate come

$$I_{m1} = V_0 \frac{R_2 + R_3}{R_1 R_2 + R_1 R_3 + R_2 R_3} \quad (8)$$

$$I_{m2} = V_0 \frac{R_2}{R_1 R_2 + R_1 R_3 + R_2 R_3}. \quad (9)$$

Si verifica facilmente che queste intensità di corrente di maglia danno gli stessi risultati che si ottengono con il metodo immediato riportato in precedenza. Infatti la corrente erogata dal generatore di d.d.p., che prima abbiamo indicato con I , coincide con I_1 come deve essere. Inoltre la ΔV_2 di prima si calcola in questo caso come

$$\Delta V_2 = R_2(I_{m1} - I_{m2}) = V_0 \frac{R_2 R_3}{R_1 R_2 + R_1 R_3 + R_2 R_3}, \quad (10)$$

dove il segno positivo indica che il nodo “in alto” in figura si trova a potenziale maggiore di quello “in basso”.

III. CIRCUITO UN PO’ PIÙ COMPLICATO

Consideriamo il circuito rappresentato in Fig. 2: ci sono due generatori (ideali) di d.d.p., rispettivamente di valore V_1 e V_2 , e cinque resistori, di resistenza R_{1-5} . Di questo circuito riusciremo a sapere tutto, però, tanto per porre una domanda semplice, immaginiamo di voler conoscere la differenza di potenziale ΔV_2 ai capi del resistore R_2 . Al solo scopo di permettere conti abbastanza

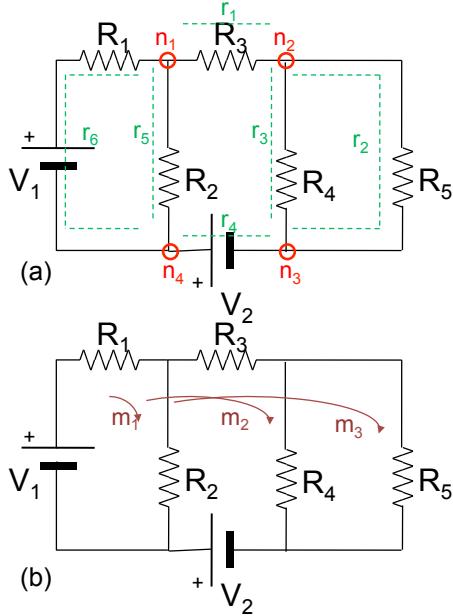


Figura 2. Schema del circuito considerato nel testo con indicati i nodi e i rami (a), e le maglie indipendenti utilizzate nella soluzione (b).

facili, immaginiamo anche $V_2 = 2V_1 = 2V$ e $R_j = jR$ (vuol dire $R_1 = R$, $R_2 = 2R$, e così via), tranne che $R_5 = R_4 = 4R$.

Cominciamo con l'individuare nodi e rami, che sono in numero $N_n = 4$ e $N_r = 6$ (tutto è indicato in Fig. 2 con la stessa notazione di Fig. 1). Il numero di maglie indipendenti è $N_m = 3$. Le tre possibili maglie m_{1-3} sono anche marcate in figura; anche qui sceglieremo come positivo il verso di percorrenza orario per le correnti di maglia.

A. Semplificazione

Prima di procedere “al buio”, conviene notare che, ai fini della domanda che ci siamo posti, il circuito può essere notevolmente semplificato. Infatti i resistori R_4 e R_5 sono evidentemente in parallelo tra loro senza nessun generatore o altro componente in mezzo, e il resistore R_3 è in serie con quel parallelo. Possiamo allora ridurre il circuito dato a quello di Fig. 3, in cui compare il resistore R_{345} che sostituisce i precedenti resistori ed equivale a quelli. Usando le regole di serie e parallelo si trova facilmente $R_{345} = 5R$.

In questo nuovo circuito i nodi sono solo due e i rami tre, per cui le maglie da considerare sono solo due.

Complessivamente nel circuito si possono individuare tre differenti maglie, quella di sinistra, quella di destra e quella esterna, come rappresentate e denominate in figura. Vedremo come, operando due diverse scelte per la scomposizione in maglie, sarà possibile giungere al-

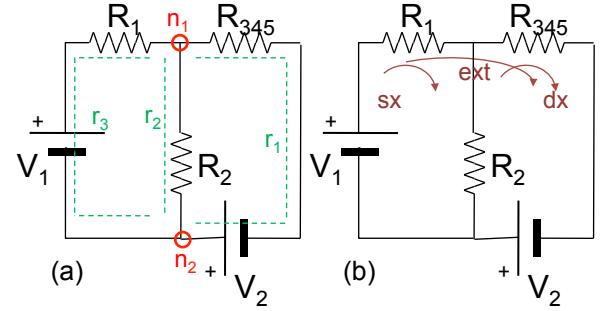


Figura 3. Schema del circuito semplificato equivalente a quello di Fig. 2 per gli scopi dell'esercizio; al solito, il pannello (a) mostra nodi e rami, e il pannello (b) le maglie considerate nella soluzione, con la denominazione usata nel testo (sx, dx ed ext stanno per sinistra, destra, esterna, rispettivamente).

lo stesso risultato. Alla fine proveremo anche un metodo differente, che è anche atteso portare allo stesso risultato.

B. Metodo 1: maglie laterali

Cominciamo considerando le due maglie laterali, di sinistra e di destra, facendo riferimento ai segni di figura per il verso positivo di circolazione delle correnti di maglia I_{m1} e I_{m2} . Notiamo poi che la resistenza R_2 è a comune fra le due maglie. Le equazioni di maglia, scritte nel modo da me preferito, sono

$$V_1 = R_1 I_{m1} + R_2 I_{m1} - R_2 I_{m2} \quad (11)$$

$$V_2 = R_2 I_{m2} - R_2 I_{m1} + R_{345} I_{m2} . \quad (12)$$

Usando le relazioni numeriche tra le varie grandezze si trova

$$V = 3RI_{m1} - 2RI_{m2} \quad (13)$$

$$2V = -2RI_{m1} + 7RI_{m2} . \quad (14)$$

Questo sistema a due equazioni algebriche lineari e due incognite può essere facilmente risolto fornendo $I_{m1} = (V/R)(11/17)$ e $I_{m2} = (V/R)(8/17)R$. Le due correnti scorrono in senso inverso attraverso il resistore R_2 . Dunque per la legge di Ohm si avrà $\Delta V_2 = R_2(I_{m1} - I_{m2}) = (6/17)V$, dove il segno positivo indica che il punto più in alto del resistore in figura si trova a potenziale maggiore.

C. Metodo 2: maglie sinistra e esterna

Vediamo cosa succede se sceglieremo quest'altra suddivisione del circuito in maglie. Indicando con I'_{m1} e I'_{m2} le nuove correnti di maglia, e osservando che stavolta è R_1 ad essere in comune tra le maglie (occhio: le correnti di maglia hanno lo stesso verso su questo resistore e quindi producono d.d.p. dello stesso segno), le equazioni

relative saranno

$$V_1 = R_1 I'_{m1} + R_1 I'_{m2} + R_2 I'_{m1} \quad (15)$$

$$V_1 + V_2 = R_1 I'_{m2} + R_1 I'_{m1} + R_{345} I'_{m2}, \quad (16)$$

dove abbiamo tenuto conto del fatto che i due generatori sono in serie tra loro nella maglia esterna (ancora principio di sovrapposizione delle tensioni, oppure Eq. 2).

Semplificando le espressioni come prima, si trova

$$V = 3R I'_{m1} + R I'_{m2} \quad (17)$$

$$3V = R I'_{m1} + 6R I'_{m2}. \quad (18)$$

Risolvendo si ottiene $I'_{m1} = (V/R)(3/17)$ e $I'_{m2} = (V/R)(8/17)$. Stavolta, il resistore R_2 è attraversato dalla sola corrente I'_{m1} , per cui $\Delta V_2 = R_2 I'_{m1} = (6/17)V$, che è risultato identico a quello di prima.

D. Metodo 3: sovrapposizione dei generatori

Utilizziamo qui il metodo che consiste nello scrivere equazioni in cui, di volta in volta, si considera un solo generatore di d.d.p., essendo gli altri (l'altro, in questo caso) sostituito da un corto-circuito. Risolto il sistema delle due equazioni così scritte, potremo usare il principio

di sovrapposizione per ottenere le intensità di corrente e le d.d.p. come somme algebriche di quelle ottenute da ogni equazione.

Dunque riprendiamo l'intero circuito e rimpiazziamo il generatore 2 con un cortocircuito. Quello che resta è facile da descrivere. Al generatore superstite, quello che eroga la d.d.p. V_1 , è collegata la serie di R_1 con il parallelo tra R_2 e R_{345} , con resistenza equivalente $R_{2345} = (10/7)R$. Con un po' di applicazione delle regole dei partitori di corrente e tensione si trova che la d.d.p. ai capi del resistore R_2 è $\Delta V_{2,G1} = (10/17)V$.

Facciamo la stessa operazione sostituendo il generatore 1 con un cortocircuito: stavolta al generatore che eroga la d.d.p. V_2 resterà collegata la serie di R_{345} con il parallelo di R_1 e R_2 , che ha resistenza equivalente $R_{12} = (2/3)R$. Ragionando come sopra, si trova $\Delta V_{2,G2} = -(4/17)V$, dove il segno negativo indica che in questo caso il nodo "in alto" di figura del resistore R_2 si trova a potenziale minore.

Per il principio di sovrapposizione è $\Delta V_2 = \Delta V_{2,G1} + \Delta V_{2,G2} = (6/17)V$, che è anche stavolta lo stesso risultato ottenuto prima.

Siete invitati a provare come esercizio altre possibilità (scelta di diverse maglie, trattazione del circuito originale) e anche a calcolare altre grandezze relative al circuito considerato o a inventarvi (e risolvere) altri circuiti.

[1] I circuiti considerati qui sono alimentati in continua, dato che i generatori considerati sono continui, cioè operano in condizioni stazionarie. Nel seguito vedremo un'estensione piuttosto immediata a situazioni di d.d.p. e corrente alternate.

[2] Volendo, la necessità di considerare tutte le diverse cor-

renti di maglia che interessano un ramo è conseguenza del principio di sovrapposizione, che noi abbiamo citato come ingrediente da usare nel caso di presenza di più generatori, ma che naturalmente vale "a prescindere" per tutte le grandezze fisiche di nostro interesse (in questo caso per determinare la d.d.p. ai capi di R_2).

Oscilloscopio: primo impatto

francesco.fuso@unipi.it

(Dated: version 6 - FF, 8 novembre 2019)

Questa breve nota vorrebbe aiutarvi ad avere un primo impatto sereno, e non traumatico, con l'oscilloscopio, uno strumento tanto diffuso e potente quanto ricco di funzioni e, almeno inizialmente, un po' complicato da usare. In essa si fa riferimento specifico, specie per la descrizione dei vari comandi, al modello di oscilloscopio analogico Isotech ISR-6051, o equivalente, che è attualmente presente in tutte le postazioni di laboratorio didattico. Naturalmente la lettura di questa nota non esaurisce in alcun modo la descrizione delle molteplici operazioni che un oscilloscopio è in grado di compiere, obiettivo che potrete conseguire solo con un assiduo esercizio.

I. ASPETTI GENERALI

L'oscilloscopio (o oscillografo) è senz'altro uno strumento che, ancora più di altri, va usato più con la testa che con le mani. La ricchezza di funzioni che ogni oscilloscopio possiede rende difficile capire tutto di primo acchito ed è sicuramente necessario un periodo di training per evitare di cadere in trabocchetti e impiegare in modo corretto lo strumento in funzione delle necessità della misura. La lettura del manuale, disponibile nel sito di e-learning, è utile, ma spesso non basta per avere un'idea chiara di cosa lo strumento sta facendo o di come debba essere impostato per l'analisi richiesta. In questa nota si fa riferimento al modello in uso in laboratorio (Isotech ISR-6051, o equivalente): si tratta di uno strumento analogico (oggi è molto frequente trovare oscilloscopi digitali, che hanno un principio di funzionamento diverso, ma un modo di operazione simile) a due canali (due ingressi, CH1, CH2) con banda passante di 50 MHz. Nel futuro vi troverete probabilmente a usare altri modelli: tutti si comportano in maniera simile e hanno comandi analoghi; addirittura il layout dei comandi, cioè la disposizione di manopole e pulsanti sul pannello raggruppati per funzioni, è spesso molto simile.

Alcune norme generali da seguire, più o meno banali, sono elencate qui di seguito.

- Gli ingressi dell'oscilloscopio, cioè i canali, sono fatti per ricevere, o misurare, delle *differenze di potenziale* (quindi non intensità di corrente, o altro). Tipicamente la resistenza di ingresso di questi canali è 1 Mohm, eventualmente abbassabile ponendo una resistenza esterna in parallelo (tipicamente da 50 ohm, lo vedrete al prossimo anno!).
- Gli ingressi sono *riferiti a massa*, cioè la differenza di potenziale è sempre misurata rispetto alla massa o terra del circuito. La massa è la scatola metallica dell'oscilloscopio, a sua volta collegata alla terra dell'impianto elettrico tramite il cordone di alimentazione. L'accesso elettrico alla massa è consentito da una boccola metallica sul frontale (accanto c'è il simbolino della massa).
- In conseguenza del punto precedente, per eseguire misure con l'oscilloscopio occorre assicurarsi che il punto del circuito sotto analisi che si vuole usare come riferimento per il potenziale sia collegato alla massa dell'oscilloscopio, o alla terra dell'impianto elettrico. State attenti: ci sono due banane di ingresso all'oscilloscopio, ma lo strumento *non* misura la d.d.p. tra queste due banane (che si riferiscono ai due canali di cui lo strumento è dotato)!
- Piccola nota: il problema di cui sopra è accentuato nel nostro laboratorio, che è il paese delle banane. In realtà i connettori di ingresso degli oscilloscopi sono "a due poli" (segnaletica e massa), essendo connettori coassiali di tipo BNC normalmente collegati a cavi coassiali che portano sia segnale che massa (ne ripareremo in seguito).
- Anche senza alcun collegamento a un circuito, scoprirete che l'oscilloscopio "vede" sempre, o quasi, qualcosa. Infatti la banda passante dello strumento gli permette di essere sensibile ai campi a radiofrequenza che sono normalmente nell'ambiente (i cavi di ingresso, le banane, l'eventuale vostra manina e dunque il vostro corpo sono delle belle antenne che captano questa radiofrequenza). Questi segnali fanno parte di quello che generalmente si chiama *rumore* e si ritrovano, almeno in parte, anche quando si fanno delle misure, specie se ad "alta frequenza". Altri rumori di ampiezza ancora più rilevante e difficilmente eliminabili sono quelli a 50 Hz, o armoeniche, legati all'oscillazione periodica della corrente alternata che fluisce nell'impianto elettrico. Questa corrente crea dei campi magnetici oscillanti, che si "accoppiano" facilmente ai fili in ingresso allo strumento di misura (questo tipo di rumore si chiama spesso *rumore di pick-up*).
- In conseguenza del punto precedente, per garantirsi che all'ingresso dell'oscilloscopio non sia presente alcun segnale occorre collegare l'ingresso stesso a massa. Questo può essere fatto premendo il tasto **GND** del canale, o dei canali, di interesse, che esegue questa connessione internamente allo strumento.
- L'operazione di cui sopra è rilevante per il seguente motivo. Come descriveremo in seguito, l'oscilloscopio permette di fare una rappresentazione di

un segnale (tipicamente dipendente dal tempo) attraverso un grafico, che è quello che si vede sullo schermo. Come ogni grafico, occorre determinare l'origine, cioè lo zero. Per quanto riguarda l'asse verticale, la determinazione dello zero deve essere fatta *separatamente* dalle misure, in genere *prima* di queste, ponendo l'ingresso a massa o terra (vedi sopra) e muovendo la manopola POSITION del canale, o dei canali, in uso finché la traccia non coincide con una posizione da voi prescelta della graticola che appare sullo schermo. La linea orizzontale corrispondente a questa posizione sarà lo zero della vostra misura di d.d.p..

- Analogamente, il fattore di scala del grafico deve essere aggiustato secondo necessità. A questo scopo, almeno per quanto riguarda l'asse verticale, l'operazione si compie agendo sulle manopole (a scatti) VOLT/DIV, dove div sta per *divisione*, che sarebbe il lato di un quadretto della graticola riportata sullo schermo (tipicamente di lato 1 cm).
- Occhio: le regolazioni prescelte sono specificate mediante apposite scritte che compaiono sullo schermo (in alto o in basso). Identificatele per bene: dovete arrivare al punto di saper interpretare *tutto* quello che sta scritto sullo schermo! In modo simile, dovete capire tutto quello che sta scritto sul pannello accanto ai vari tasti e manopole, senza dubbi e incertezze. Inoltre state attenti alle spie del pannello stesso, che non stanno lì a far nulla, ma servono per segnalare qualcosa. In particolare le lucine rosse che talvolta si accendono vogliono dire che dovete fare attenzione alla scelta dei parametri di funzionamento che state utilizzando.
- L'oscilloscopio, come descriveremo meglio in seguito, serve soprattutto per *visualizzare* dei segnali (d.d.p.) dipendenti dal tempo, cioè verificare la loro forma (detta anche gergalmente *forma d'onda*). Tuttavia esso viene spesso impiegato per fare misure (di d.d.p. o di intervalli temporali, per esempio). La precisione delle misure è generalmente piuttosto bassa: negli strumenti in uso in laboratorio, l'incertezza di calibrazione è, per la maggior parte delle portate di tensione e tempo (si veda il manuale per ulteriori informazioni) del 3% e ad essa va sommata ((in genere)convenzionalmente in quadratura) l'incertezza di lettura dovuta allo spessore finito della traccia sullo schermo, un po' come quando si fanno misure da un grafico disegnato con una matita non particolarmente fine. Dunque è normale che le misure fatte con un oscilloscopio abbiano barre di errore non trascurabili.
- L'oscilloscopio è uno di quegli strumenti che non amano essere spenti spesso. Esso infatti deve "termalizzare" affinché le tolleranze di calibrazione siano rispettate. Ricordatevene, ma ricordate anche di evitare che lo strumento resti acceso con un bel

punto luminoso fisso sullo schermo, come si può ottenere operando in modalità X-Y, o Y-X (vedremo poi di cosa si tratta) senza collegare niente agli ingressi. Nel caso, riducete la luminosità dello schermo agendo sull'apposita manopola INTENS.

- Cosa ovvia, ma va detta: è severissimamente vietato giocherellare con calamite e magneti dalle parti dello schermo dell'oscilloscopio, che potrebbe venire danneggiato in modo irreparabile!

II. PRINCIPIO DI FUNZIONAMENTO

Il principio di funzionamento dell'oscilloscopio può essere compreso facendo riferimento a una descrizione che probabilmente avete conosciuto alle scuole superiori, o che potete trovare facilmente in rete con tanto di belle figurine. Questa descrizione semplice semplice si attaglia perfettamente alla descrizione degli oscilloscopi analogici come quello che usate quest'anno. In sostanza, l'obiettivo della descrizione è mostrare che quello che si vede sullo schermo è un grafico prodotto da una penna che può muoversi molto rapidamente e usa un inchiostro che si cancella dopo un tempo breve, paragonabile al tempo tipico di persistenza delle immagini sulla retina umana.

Elemento chiave dell'oscilloscopio analogico è il tubo a raggi catodici. Questo non è altro che un tubo di vetro in cui è stato fatto il vuoto e di cui una faccia è lo schermo. All'interno si trova una sorgente di elettroni costituita da un filamento riscaldato per effetto Joule e fatto di un materiale (tungsteno) in grado di liberare elettroni per effetto termoionico. Questi elettroni vengono accelerati da una d.d.p. dell'ordine della decina di kV in modo da arrivare con grande energia cinetica sullo schermo sotto forma di un fascio molto collimato (di sezione *idealmente* "puntiforme"), detto *pennello elettronico*, che ha il ruolo della penna ipotizzata prima. Poiché la penna è fatta di elettroni, che hanno una massa molto piccolo, è chiaro che essa può spostarsi con grande velocità, ovvero subire forti accelerazioni. Lo schermo è ricoperto di un materiale a base di fosforo che si illumina dove è colpito dal pennello elettronico. Inoltre il meccanismo di produzione della luce (fosforescenza) fa sì che l'illuminazione resti attiva per un certo tempo (*persistenza*), per cui se il pennello si sposta sullo schermo l'occhio può, in certe condizioni, vedere una *traccia*, cioè una linea continua, pur in presenza di spostamenti molto rapidi. La fosforescenza permette quindi di individuare l'inchiostro che si era ipotizzato prima.

All'interno del tubo a raggi catodici sono collocate due coppie di placchette metalliche collegate all'esterno del tubo. Queste coppie di placchette formano dei campi elettrici (in prima approssimazione come condensatori ad armature piane e parallele) il cui scopo è quello di *deflettere* il pennello elettronico. Infatti esse sono disposte in modo che il pennello ci passi attraverso; la geometria, poi, è scelta in modo da provocare deflessioni in due

direzioni ortogonali, che chiameremo orizzontale e verticale, oppure X e Y. Queste sono proprio le direzioni della rappresentazione grafica che si vede sullo schermo dell'oscilloscopio.

La Fig. 1, tratta da <http://www.hit.bme.hu/~papay/edu/Ana/Scope.htm>, riporta uno schema di massima del tubo catodico e dà indicazioni sul suo funzionamento quando l'oscilloscopio è operato in modalità Y-t (poi vedremo di cosa si tratta).

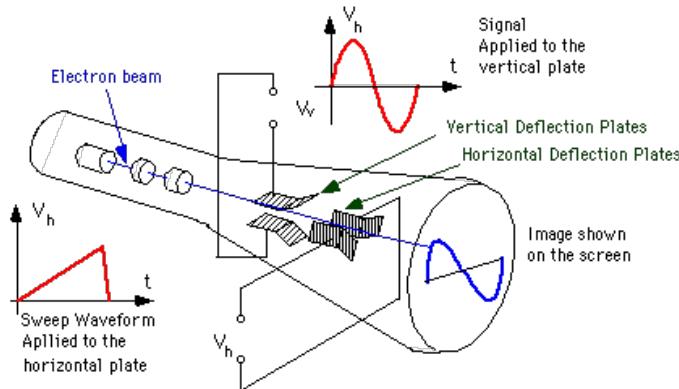


Figura 1. Schema di massima del tubo catodico.

Grazie a diversi accorgimenti costruttivi e progettuali, si fa in modo che la deflessione sia proporzionale all'intensità del campo elettrico generato dalle coppie di placchette, cioè alla *differenza di potenziale* che vi è applicata. Dunque la posizione di impatto del pennello elettronico sullo schermo, cioè la posizione del punto illuminato, dipende in maniera pressoché proporzionale dalla coppia di differenze di potenziale applicata sulle placchette di deflessione lungo X e lungo Y. Fate attenzione: affinché il pennello possa essere deflesso "correttamente", cioè in modo che la sua posizione sullo schermo sia controllata linearmente dalla d.d.p. applicata tra le placchette, occorrono alcuni accorgimenti tecnici sul dimensionamento delle placchette stesse. Pensateci su usando le vostre conoscenze di elettrostatica.

A. Modo di operazione X-Y

In prima battuta, l'oscilloscopio è un aggeggi che permette di visualizzare su uno schermo un punto le cui coordinate sono proporzionali a due tensioni (entrambe riferite a massa, ricordate!). Questa è effettivamente una modalità di operazione dell'oscilloscopio, cosiddetta X-Y (o anche Y-X), in cui canale 1 e canale 2 dell'oscilloscopio accettano la coppia di tensioni che rappresentano, a meno di fattore di scala e origine, le coordinate del punto che si ottiene sullo schermo. Ad essa si accede premendo il tasto X-Y, la cui azione è segnalata da un'indicazione in basso a destra dello schermo. Origine e fattore di scala del grafico possono essere variati agendo sui comandi

dello *stadio di ingresso* secondo quanto specificato nella sezione successiva.

L'utilità di avere uno strumento che grafica una coppia di tensioni vi sarà chiara con l'uso. Qui anticipiamo qualcosa, partendo dalla constatazione che analizzare in questo modo due d.d.p. costanti (due *segnali continui*) non è molto eccitante, dato che si ottiene un punto. Però, come sapete, il mondo è fatto di *segnali variabili*, o segnali tout-court, cioè di tensioni che variano nel tempo, e un caso particolarmente interessante, anche se non esauritivo, è quello dei segnali che dipendono armonicamente dal tempo. Vi ricordate di sicuro come si scrivono le proiezioni sugli assi cartesiani del moto di un punto che percorre una traiettoria circolare: si tratta di due funzioni posizione che vanno con il tempo come il seno e il coseno, cioè di due funzioni "armoniche" (il termine non è molto corretto, si intende qui tipo seno o coseno) *sfasate* nel tempo tra loro di un termine costante pari a $\pi/2$. Non ci vuole molto a rendersi conto che, se lo sfasamento è a un angolo diverso, allora la circonferenza di defessa in un'ellisse, in cui l'orientazione degli assi minore e maggiore, misurata rispetto alle direzioni cartesiane, e l'ellitticità dipendono proprio dallo sfasamento. Analogamente non è difficile rendersi conto che se lo sfasamento non c'è, o vale π (o suoi multipli interi), allora l'ellisse collissa in una retta.

Spesso in laboratorio si ha la necessità di misurare, o stimare, lo sfasamento tra due segnali periodici alla stessa frequenza (e coerenti fra loro). In questi casi l'uso dell'oscilloscopio in modalità X-Y permette di verificare in modo immediato la presenza di uno sfasamento, come vedrete in qualche esercitazione pratica.

III. STADIO DI INGRESSO

Lo stadio di ingresso è quello che "condiziona" (manipola elettronicamente) le tensioni in ingresso a CH1, CH2 prima di mandarle alle placchette di deflessione. Le funzioni dello stadio di ingresso sono molteplici. Qui mi limiterò a segnalare le più importanti, che riguardano non solo la modalità X-Y ma anche quella Y-t di cui tratteremo in seguito.

- Fattore di scala.** Modificare il fattore di scala della rappresentazione significa agire sul "guadagno" (cioè amplificazione o attenuazione) del segnale in ingresso. Il fattore di scala si modifica con la manopola a scatti VOLT/DIV (ogni canale ne ha una), che è calibrata in V per divisione (lato del quadretto della graticola). La sensibilità massima è tipicamente dell'ordine di qualche mV/div e le divisioni sono una decina. Occhio: se il fattore di scala è sbagliato, ad esempio eccessivamente alto o basso, allora può essere che il punto disegnato dal pennello elettronico vada fuori schermo o sembra non muoversi dal punto che avete stabilito come origine. Dunque, se "non vedete nulla", pensate all'eventua-

lità che il fattore di scala (e anche l'offset, di cui tratteremo fra poco) sia aggiustato male.

- **Scalibrazione.** Normalmente il fattore di scala è calibrato entro la precisione fornita dal costruttore. Il fattore di scala si legge sullo schermo (ogni canale ha la propria indicazione!). In certi casi può essere utile “scalibrare” la scala, ad esempio per ottimizzare la “dinamica di rappresentazione”. Per scalibrare, nel modello di oscilloscopio preso qui a riferimento occorre premere sulla manopola della scala: si accende una bella spia rossa e ruotando la manopola si può modificare la scalibrazione. Al posto del fattore di scala, ovvero della sensibilità verticale in V/div, sullo schermo compare un punto interrogativo, dato che non è più possibile conoscere il fattore di scala stesso (è scalibrato!).
- **Offset di scala.** L'origine del grafico disegnato dall'oscilloscopio si trova in una posizione che è stabilita, tecnicamente, aggiungendo una tensione costante (*offset*) a quella che si vuole visualizzare. L'operazione si compie agendo sulle manopole (non a scatti e un po' più piccole di quelle del fattore di scala) che si chiamano **POSITION**. Ce ne è una per canale e inoltre un'altra per la posizione orizzontale. La posizione dell'origine si determina rispetto a un punto (o una retta orizzontale, vedi dopo per la modalità Y-t) della graticola sullo schermo. Quando si determina la posizione dell'origine è necessario assicurarsi che non ci siano segnali in ingresso, cioè è necessario premere il pulsante **GND** (sta per ground) del canale corrispondente. Questo fa comparire un simbolino di terra, o di massa, sullo schermo, vicino all'indicazione della sensibilità verticale. Occhio! Solo dopo aver fatto questa operazione (e *solo se l'accoppiamento di ingresso è DC*, vedi dopo) è possibile stabilire il valore della tensione (riferita a massa) che si sta misurando, che è rappresentata dalla distanza, positiva o negativa, in divisioni o frazioni di divisione, del punto, o linea, che si osserva rispetto al punto, o linea, che avete stabilito essere l'origine.
- **Accoppiamento del canale di ingresso.** Sopra al pulsante **GND** ce ne è uno con scritto **AC/DC**. Non si tratta del nome di un noto gruppo musicale, da me non particolarmente amato, ma di un pulsante che cambia il comportamento dello stadio di ingresso. In DC (compare un segno = sullo schermo in basso a sinistra accanto all'indicazione della sensibilità verticale in V/div) l'ingresso del canale è accoppiato direttamente alle placchette di deflessione (DC significa direct current, cioè corrente continua). In AC si interpone un filtro (scopriremo essere un filtro “passa alto” che ha una “frequenza di taglio” molto bassa, tipicamente dell'ordine dell'1Hz) che taglia, cioè elimina, la componente continua, cioè che non varia nel tempo, del segnale

che si sta osservando. Nella pratica, questo filtro è realizzato interponendo un condensatore fra segnale e ingresso dell'oscilloscopio. Il condensatore, infatti, blocca la componente costante nel tempo del segnale. La possibilità di accoppiamento AC è estremamente utile, dato che spesso i segnali sono fatti da una parte continua e da una oscillante sovrapposta a questa. Se l'interesse è nella componente oscillante, è inutile dare peso a quella continua. Ciò permette di ottimizzare in maniera molto semplice la sensibilità della misura. D'altra parte la presenza del condensatore in certe condizioni “distorce” il segnale, attraverso un'operazione che, al limite, assomiglia a fare la “derivata temporale” del segnale. Di conseguenza quando l'accoppiamento del canale di ingresso è AC, la scala verticale non rappresenta più (necessariamente) l'ampiezza della d.d.p. applicata, ma può rappresentare l'ampiezza della sua “derivata temporale”.

IV. MODO DI OPERAZIONE Y-T

Come dice il suo nome, di norma l'oscilloscopio è usato per visualizzare dei segnali oscillanti, cioè dipendenti dal tempo (transienti o segnali periodici che siano). La capacità di visualizzare transienti veloci dipende in primis dalla banda passante dello strumento. Se questa vale 50 MHz, come nel modello di riferimento, allora si può essere confidenti di poter visualizzare transienti su una scala di poche centinaia di nanosecondi (nelle nostre esperienze non arriveremo, purtroppo, ad analizzare segnali così “veloci”, cosa che richiede parecchie ulteriori cure sperimentali).

A. La sweep

Il metodo per ottenere la visualizzazione di un segnale dipendente dal tempo è banale. Basta pilotare lo spostamento orizzontale del pennello elettronico con una tensione che dipende linearmente dal tempo. Tenendo conto della persistenza sullo schermo e del tempo di risposta del nostro occhio, questo permette direttamente di visualizzare una traccia che rappresenta l'andamento di un segnale in ingresso a uno, o tutti e due (vedi dopo), i canali dell'oscilloscopio.

Tecnicamente l'operazione è possibile grazie alla presenza all'interno dell'oscilloscopio di un generatore di tensione linearmente dipendente dal tempo. Questo generatore produce un'onda a *dente di sega*, con una crescita lineare nel tempo e una decrescita molto più rapida (il tempo necessario perché il pennello elettronico si riposiziona al punto di partenza, tipicamente a sinistra dello schermo - il tempo scorre sempre da sinistra a destra). Il movimento di spazzata orizzontale del pennello elettronico si chiama *sweep* e la velocità di spazzata, che de-

ve essere chiaramente adattata al segnale che si intende visualizzare, si chiama in genere *base dei tempi*.

Per regolare la velocità di spazzata si agisce sulla manopola a scatti TIME/DIV: la scala prescelta è scritta sullo schermo e anche qui è possibile scalibrare (premendo la manopola, si accende una spia rossa) e modificare la posizione di inizio (manopola POSITION posizionata sopra alla TIME/DIV). Tanto per dare un numero, considerate che nel modello di riferimento è possibile scegliere una base dei tempi che corrisponde al minimo a 20 ns/div: una divisione è lunga circa 1 cm, per cui la velocità orizzontale del pennello elettronico corrisponde a qualcosa dell'ordine di 5×10^5 cm/s (sono elettroni, è facile farli viaggiare rapidamente!). Ovviamente la base dei tempi va regolata sulla base della rapidità con cui varia il segnale che si sta osservando, ed è bene avere prima un'idea di quello che si vuole visualizzare per evitare di non osservare correttamente il segnale di interesse.

Come già più volte affermato, il giochino dell'oscilloscopio consiste nel permettere all'occhio di osservare la traccia lasciata dal pennello sullo schermo anche quando questa traccia è scritta molto velocemente. Questo si ottiene grazie alla fosforescenza e al fenomeno di persistenza dell'immagine sulla retina. Per permettere la cancellazione della traccia dallo schermo e dalla retina, fra una spazzata e la successiva deve intercorrere un *tempo morto*. Esso vale normalmente una piccola frazione di secondo e può essere regolato entro certi limiti (comando HOLDOFF, si veda in seguito). Poiché sia il comportamento del fosforo che quello della retina dell'occhio umano sono influenzati da numerosi fattori, soggettivi ed oggettivi, qualche volta può verificarsi che più di una traccia compaia sullo schermo. Normalmente il problema si risolve usando in modo corretto le funzionalità del circuito di trigger (si veda in seguito).

B. Canali

Gli strumenti in dotazione hanno la possibilità di visualizzare due canali indipendenti. Questo tornerà utilissimo quando dovete confrontare gli andamenti di due distinti segnali (due tensioni riferite alla stessa massa), tuttavia spesso avrete da analizzare un solo segnale. La visualizzazione dei canali si attiva o disattiva (spie continue verdi accese o spente) premendo i pulsanti CH1 e CH2. Se dovete vedere solo un canale, per esempio CH1, dovete accertarvi che questo canale sia attivo e, operazione consigliabile per evitare confusione, disattivare l'altro. Ovviamente i due canali possono avere regolazioni del tutto indipendenti per offset e scala. Le scale sono tutte e due scritte sullo schermo (attenzione a non fare confusione tra i canali!).

La visualizzazione "simultanea" dei due canali può avvenire con diverse modalità. Essa può essere *alternata*, cioè una sweep per un canale e la successiva per l'altro (utile per scelte dei basi di tempi molto brevi) oppure *chopped*. In questo caso la sweep viene suddivisa in tanti

piccoli sottointervalli temporali in cui la visualizzazione dei due canali viene alternata. In genere, grazie alla persistenza della traccia, si ottiene la sensazione di vedere contemporaneamente i due canali (il chopping risulta evidente operando a valori grandi della base dei tempi). Il passaggio da una modalità all'altra si esegue premendo i pulsanti ALT o CHOP (quest'ultima dovrebbe essere il default), da localizzare sul pannello. Ovviamente lo schermo indica la modalità prescelta (ma non sempre, la modalità chopped, di default, è spesso non indicata) e, altrettanto normalmente, l'oscilloscopio decide da solo quale impiegare in funzione della base dei tempi impostata, ma la scelta da lui compiuta può talvolta risultare poco adeguata. Ci sono poi altre modalità, per esempio quella consistente nella visualizzazione di un segnale *somma* (SUM) dei due canali, ma in genere se ne fa poco uso. State attenti a non premere inavvertitamente il pulsante corrispondente, che fa comparire un segno "+" sullo schermo tra le indicazioni di scala dei due canali. Analogamente state attenti a non attivare il comando (INV) che inverte il segnale del canale CH2, utile ad esempio per passare dalla somma di cui prima alla differenza tra i segnali dei due canali.

V. TRIGGER

La parte forse più rognosa nell'uso dell'oscilloscopio è quella che riguarda l'uso del trigger. Per capire l'utilità delle funzioni che il trigger offre pensate di dover visualizzare un transiente che si verifica ogni tanto, per esempio la scarica di un condensatore come in una delle vostre esercitazioni pratiche.

Per avere una visualizzazione corretta e completa del transiente è necessario che la partenza della sweep sia sincronizzata in qualche modo con l'evento che state osservando. Notate che il problema della sincronizzazione si presenta anche nel caso di un segnale periodico: se non c'è sincronizzazione tra la partenza della sweep e il segnale stesso, la visualizzazione risulterà probabilmente poco intellegibile, perché la forma d'onda visualizzata vi apparirà "in movimento", o addirittura, come già preannunciate, potreste vedere sullo schermo una sovrapposizione confusa di tante forme d'onda, cosa che è assolutamente da evitare.

Il circuito di trigger (vuol dire grilletto) di cui sono dotati gli oscilloscopi provvede proprio alla sincronizzazione. In modalità AUTO la sweep è, generalmente, sempre attiva, cosa utile perché permette ad esempio di regolare il livello di zero, ma cosa inutile se si vuole sincronizzare (in realtà anche in questa modalità l'oscilloscopio prova da solo a sincronizzarsi con il segnale, senza però necessariamente riuscirci...). In modalità normale (NML, un tasto provvede al passaggio da una modalità all'altra) il trigger può essere controllato dall'utente in maniera quasi completa, cioè le condizioni per la partenza della sweep possono essere determinate. Il passaggio tra le due modalità è segnalato dalle spie verdi ATO e NML in alto a

destra sul pannello. L'attivazione del trigger è segnalata da un'ulteriore spia verde posta in prossimità delle altre (subito sopra di esse).

In modalità normale si può regolare il *livello di trigger* agendo sulla manopola **LEVEL**: si tratta del valore di tensione che, se viene raggiunto (a salire o a scendere, a seconda della **SLOPE** che può essere commutata da positiva a negativa agendo sul tasto omonimo - osservate quale indicazione cambia sullo schermo, è un simbolino di transiente con una freccina) dal segnale in ingresso, fa partire la sweep, che altrimenti resta spenta. Affinché la sincronizzazione avvenga con il segnale di vostro interesse occorre naturalmente dire al circuito di trigger quale ingresso deve monitorare: agendo sul tasto **SOURCE** si può scegliere tra **CH1**, **CH2**, **EXT** (un ulteriore canale, accessibile con un connettore BNC dedicato, non visualizzabile ma utilizzabile solo come trigger), **VERT** (il trigger monitora alternativamente i due canali e fa partire la sweep alternativamente rispetto al livello dell'uno e dell'altro, scelta specifica per alcune operazioni particolari e normalmente da evitare), **LINE** (in questo caso la sincronizzazione avviene rispetto alla corrente alternata a 50 Hz dell'impianto elettrico - utile alcune volte, come vedrete). Ci sono poi altre possibilità, per esempio quelle accessibili con il tasto **TV** e poi qualche altra specificazione, che servono per i riparatori di televisori, if any, e che quindi *non dovete mai usare*.

Un'ulteriore vantaggiosa complicazione è data dal fatto che è possibile interporre tra ingresso monitorato e circuito di trigger dei filtri (**AC**, **HFR**, **LFR**, che si commutano con il tasto **COUPLING**). I filtri HFR e LFR attenuano rispettivamente le componenti ad alta e bassa frequenza del segnale monitorato ai fini del trigger. Per la spiegazione dell'utilità rimando senz'altro a qualche esempio pratico.

Osservate in generale che, in accordo con il manuale, la regolazione del livello di trigger è per sua natura *approssimativa* e infatti essa non è quantificata sullo schermo. Inoltre, il trigger scatta solo in presenza di transienti (non è possibile "accoppiarlo" in DC, almeno per il modello in uso), per cui se si vogliono visualizzare segnali continui è necessario in genere usare la modalità automatica di trigger. Comunque, a parte il livello, tutte le altre regolazioni del trigger sono annunciate con apposite scritte sullo schermo, ragione in più per guardare non bene, ma benissimo, tutto quello che vi compare!

Anche se l'uso del trigger richiede esperienza, e l'esperienza si fa solo usando lo strumento nelle più diverse condizioni, siete caldissimamente invitati a provare fin da subito a passare dalla modalità **AUTO** (in genere quella di default all'accensione) alla **NML**, per poi vedere cosa succede e ragionarci sopra.

VI. MISURARE CON L'OSCILLOSCOPIO

Fatte salve tutte le considerazioni prima riportate su precisione e risoluzione, i cui dati vanno controllati sul

manuale, l'oscilloscopio può essere tranquillamente usato per fare misure, e questo voi farete durante l'anno (questo e anche il prossimo). Le misure sull'oscilloscopio si fanno come le misure su un grafico disegnato su carta millimetrata. Si stimano le distanze in divisioni (o frazioni di divisione, la divisione è suddivisa in cinque tacchette, guardate lo schermo) tra punti "interessanti" del segnale (picchi, minimi, massimi, etc.) e poi, usando il fattore di scala, si deducono i valori degli intervalli misurati (di tempo o di tensione).

Per le nuove generazioni, però, si è pensato a un aiuto costituito da una coppia di *cursori* (in realtà si tratta di linee tratteggiate, orizzontali o verticali a seconda che si vogliano fare misure di grandezze sulla scala rispettivamente verticale o orizzontale, cioè generalmente differenze di potenziale o di tempi) che possono essere posizionati agendo sulla manopola **VARIABLE**. I tasti accanto a questa manopola gestiscono i cursori e il loro movimento secondo una logica che può essere facilmente capita leggendo cosa c'è scritto sul pannello accanto ai vari pulsanti e, soprattutto, facendo prove. Notate che premendo la manopola si passa da un movimento fine a un movimento grossolano e che si può spostare un cursore alla volta (o tutti e due assieme), essendo data la possibilità di scegliere quale cursore muovere agendo, appunto, su un pulsante. Il cursore che si può muovere è segnalato da una freccina. Normalmente lo step minimo con cui potete muovere un cursore è 1/25 della divisione. Fate anche attenzione al fatto che, spesso, lo spessore della traccia non è affatto trascurabile, circostanza che dovrebbe indurvi a esprimere sempre in modo corretto l'incertezza di misura. Infatti l'indicazione relativa ai cursori, che compare in alto a sinistra sullo schermo, *non tiene conto* di alcuna incertezza, né di calibrazione né di lettura.

L'uso dei cursori è molto più semplice nella pratica che non nella descrizione. La distanza tra i cursori, nelle unità corrette (e debitamente scalate), è riportata sullo schermo in alto a sinistra. Nel caso si misurino dei periodi, è possibile visualizzare direttamente la misura della frequenza, cioè il suo reciproco. Fate la massima attenzione, però: se state visualizzando due tracce, ovvero due canali contemporaneamente, come specificato sopra, controllate sempre a quale canale fa riferimento la misura dei cursori. Infatti nelle misure di d.d.p. i cursori si riferiscono al canale CH1 (sullo schermo leggete ΔV_1); per leggere il valore nella scala di CH2 occorre spegnere CH1.

Detto che il mio consiglio, almeno all'inizio e quando si può, è di non fare uso dei cursori ma affidarsi alle tacchette, vi segnalo che lo schermo dell'oscilloscopio riporta pure una misura di frequenza ($f =$ e poi un numero con tante cifre significative in basso a destra). Questa misura non ha necessariamente a che fare con quello che state visualizzando sullo schermo, essendo prodotta da un dispositivo a parte, un *frequenzimetro*, integrato nell'oscilloscopio. Questo frequenzimetro misura, o cerca di misurare, la frequenza dell'eventuale segnale periodico utilizzato come trigger, dunque osserva il canale o segnale

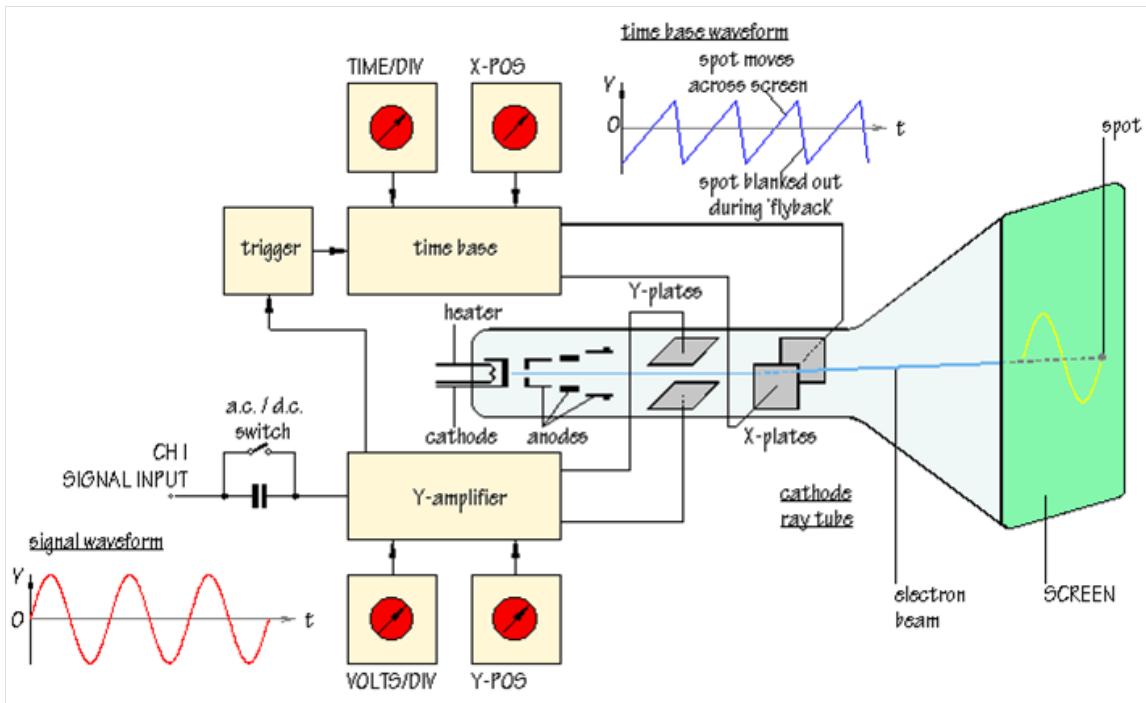


Figura 2. Uno schema a blocchi dell'oscilloscopio.

che è selezionato come sorgente del trigger. A meno che non vi si consigli diversamente, e almeno all'inizio, *fate finta che questa indicazione non ci sia*. In caso contrario è alta la probabilità di trarre conclusioni poco corrette.

VII. ALTRE

Nell'oscilloscopio ci sono molte altre funzioni e molti altri tasti. Alcune di queste funzioni le discuteremo quando avremo da risolvere dei problemi specifici, cioè da eseguire misure particolari. Di altre probabilmente non parleremo mai, non essendo utili per i nostri scopi. Siete sempre invitati a esplorare, cioè a premere qualche pulsante o girare qualche manopola e vedere cosa succede (in genere le indicazioni sul pannello o gli annunci sullo schermo sono abbastanza auto-esplicativi). Se lo fate, tenete conto che una barretta orizzontale disegnata sul pannello in prossimità di qualche tasto significa che, per attivare o disattivare la funzione corrispondente, occorre tenere premuto il tasto per qualche secondo (e si sente un beep quando il comando viene eseguito).

Ricordate sempre che alcune delle funzioni possono dare un esito che, se non debitamente considerato, può com-

promettere la validità delle vostre osservazioni. Questo è il caso, per esempio, della funzione INV, già citata, che inverte il segno del segnale in ingresso al canale 2 (e compare una freccina che punta verso il basso sullo schermo), delle funzioni MAG, che servono per espandere la scala orizzontale creando in genere confusione, della funzione P10 (o altro numero), che ha la subdola funzione di aumentare artificialmente di un fattore 10 (o altro numero) la scala verticale. Fate la massima attenzione a non attivare involontariamente queste funzioni!

VIII. SCHEMA A BLOCCHI

Navigando in rete si possono trovare numerosi schemi a blocchi dell'oscilloscopio. Io ne ho selezionato uno, il primo che mi è capitato, dal sito <http://www.doctrionics.co.uk/scope.htm>, e lo riporto in Fig. 2. Questo schema a blocchi, che fa chiaramente riferimento alla modalità Y-t, è abbastanza auto-esplicativo, pur se non molto completo: il mio consiglio è quello che voi proviate a ricostruire da soli un ragionevole schema a blocchi, cercando di collocare i vari comandi di cui abbiamo trattato all'interno dei rispettivi blocchi concettuali di pertinenza.

Carica/scarica condensatore con Arduino e onda quadra

francesco.fuso@unipi.it

(Dated: version 1 - Francesco Fuso, 1 novembre 2021)

In questa nota si descrive brevemente l'esercitazione pratica sull'acquisizione dei dati via Arduino applicata alla carica/scarica di un condensatore (e altro), mettendo in luce strategia di misura e tecniche di acquisizione e trattamento dati.

I. INTRODUZIONE

L'esercitazione pratica ha diversi scopi: il principale scopo "scientifico" è l'acquisizione di dati rilevanti per ricostruire il processo di carica e scarica di un condensatore di capacità C su una resistenza R di valore conosciuto, cioè misurato con multmetro. Come illustrato nel seguito di questa nota, la strategia di acquisizione prevede di registrare in maniera continuativa la d.d.p. ai capi del condensatore usando come d.d.p. per la carica un segnale con forma d'onda onda quadra opportunamente configurata. Inoltre l'esperimento prevede di costruire record abbastanza corposi in termini di lunghezza, con la possibile conseguenza di registrare diversi cicli di carica e scarica. I dati, debitamente trattati, potranno servire per avere una misura di C , che non può essere eseguita con i multimetri a nostra disposizione [1].

II. CONFIGURAZIONE DI MISURA

La configurazione concettuale più semplice e immediata per studiare carica e scarica di un condensatore è rappresentata in Fig. 1(a): a un dato istante $t_0 = 0$ uno switch (un commutatore, per esempio ad azionamento meccanico) viene commutato sulla posizione 1 e il condensatore C , che supponiamo precedentemente scarico, viene caricato dal generatore attraverso la resistenza R ; di conseguenza, la d.d.p. ai suoi capi cresce nel tempo secondo la funzione $V(t) = V_0[1 - \exp(-t/\tau)]$ tendendo a giungere a un valore asintoticamente pari a V_0 . A un istante successivo, che chiamiamo t' , il commutatore passa sulla posizione 2 e il condensatore, caricato "completamente" nella fase precedente, si scarica attraverso la resistenza R ; di conseguenza, la d.d.p. ai suoi capi diminuisce esponenzialmente nel tempo secondo la funzione $V(t) = V_0 \exp[-(t - t')/\tau]$, tendendo a giungere asintoticamente a zero. Nelle equazioni, supponendo il generatore ideale, il *tempo caratteristico* τ è pari a RC .

La configurazione sperimentale effettivamente impiegata è invece mostrata in Fig. 1(b): lo switch non c'è e la sua funzione è in pratica sostituita dall'uso di una forma d'onda quadra, cioè una d.d.p. che può assumere due distinti valori nel tempo [2] concettualmente corrispondenti alle due posizioni dello switch. Nel seguito indicheremo con V_0 il valore di questa d.d.p. nel semiperiodo in cui l'onda quadra è a "livello alto" e porremo pari a zero il suo valore nell'altro semiperiodo. Quest'ultima condi-

zione, che richiede di operare con attenzione sull'`offset` del generatore di funzioni e che può essere verificata solo entro l'incertezza delle misure fatte con oscilloscopio, non è strettamente necessaria, dato che influenza solo sul livello asintotico del segnale misurato, ma aiuta nella descrizione. Volendo utilizzare Arduino, che può digitalizzare d.d.p. in ingresso esclusivamente positive e di valore massimo prossimo a 5 V, un altro requisito sperimentale è $0 < V_0 \lesssim 5$ V (circa).

Il funzionamento del circuito è ovvio: nel semiperiodo in cui è presente la d.d.p. V_0 il condensatore si carica, quando invece la d.d.p. fornita dal generatore va a zero il condensatore si scarica: ricordate, infatti, che il generatore di funzioni, come ogni generatore reale, ha una sua resistenza interna (di valore $r_G = 50$ ohm, con tolleranza 10%, secondo manuale) che nella fase di scarica chiude il circuito su se stesso. Il carattere reale del generatore implica ovviamente che la costante tempo effettiva diventi $\tau = (R + r_G)C$, che può o meno essere approssimata con RC in funzione del dimensionamento di R e dell'accuratezza della misura.

Lo schema di Fig. 1(b) mostra tutte le connessioni da realizzare, comprese quelle alla porta A0 di Arduino che verrà usata come ingresso analogico al digitalizzatore. Inoltre è esplicitamente indicato il collegamento ai due canali dell'oscilloscopio, che è indispensabile per verificare il funzionamento del circuito e, soprattutto, per accertarsi che siano soddisfatti tutti i requisiti necessari (per esempio l'ampiezza di V_0 , il fatto che l'onda quadra non diventi mai negativa, ma anche, come discusso in seguito, che le scale temporali siano adeguate per il campionamento). Come sapete, l'oscilloscopio è un misuratore reale di d.d.p. e quindi ha una propria resistenza interna, $r_O = 1$ Mohm nominale, che è alta ma non altissima (tenete presente che la resistenza interna del digitalizzatore di Arduino è nominalmente pari a 100 Mohm, cioè molto maggiore). Se non ve la sentite di battezzare come trascurabili gli effetti di r_O [3] potete scollegare l'oscilloscopio nelle fasi di acquisizione e usarlo solo in fase di verifica. Infine notate nello schema un ulteriore collegamento tra una porta digitale di Arduino (pin 5) e un altro segnale prodotto dal generatore di funzioni, collegamento usato per scopi di sincronizzazione come discusso in seguito.

Ricordate che, prima di collegare fisicamente Arduino al circuito, dovete aver caricato (mediante "upload") lo sketch necessario per l'esperimento e, soprattutto, aver controllato con l'oscilloscopio che i livelli della d.d.p. prodotta dal generatore di funzioni siano compatibile con le

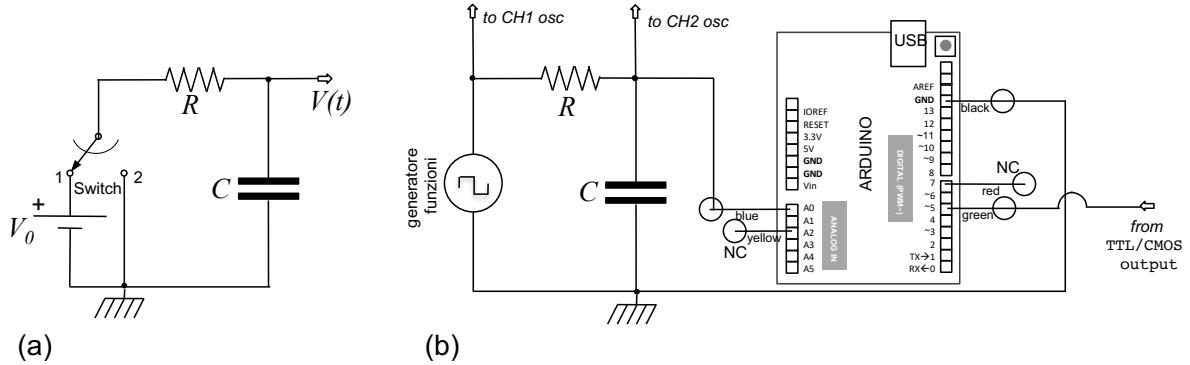


Figura 1. Schema concettuale di un esperimento di carica/scarica del condensatore (a) e sua realizzazione con Arduino come discussa nel testo (b). Nel pannello (b) è rappresentata una visione molto schematica e non in scala della scheda Arduino Uno rev. 3 SMD edition usata nell'esercitazione. Ci sono cinque collegamenti ad altrettanti pin della scheda che terminano con boccole volanti di diverso colore; secondo quanto indicato nello schema solo tre boccole devono essere collegate (NC significa non collegato). Le frecce indicate collegamenti con strumenti non mostrati in figura, per esempio con i due canali dell'oscilloscopio e con il segnale TTL/CMOS OUTPUT proveniente dal generatore di funzioni.

richieste di Arduino. Il montaggio del circuito è banale, dato che Arduino è dotato di boccole volanti colorate come da schema. Per il funzionamento del circuito è indispensabile che siano eseguiti tutti i collegamenti a massa, o terra (indicata con una spazzolina nello schema), necessari: in particolare il polo negativo del generatore deve essere collegato sia alla boccola di messa a terra dell'oscilloscopio che alla boccola GND (nera) di Arduino.

A. Sincronizzazione e campionamento

Dal punto di vista concettuale Arduino compie un’operazione banale: a intervalli temporali prestabiliti, che chiamiamo qui Δt_{samp} e che possono essere aggiustati tra 100 e 900 μs nominali agendo sullo script di Python che controlla l’esperimento (dettagli in Appendice), esso campiona la d.d.p. $V(t)$ ai capi del condensatore, digitalizzandola per renderla poi disponibile in un file trasferito al computer di laboratorio. Questo file ha due colonne di cui la seconda è proprio il valore digitalizzato (dunque in *digit*) e la prima riporta l’istante di campionamento misurato da Arduino ed espresso in unità di microsecondi. Convenzionalmente attribuiamo un’incertezza ± 1 digit alla misura di $V(t)$ e $\pm 4 \mu\text{s}$ a quella del tempo.

Normalmente l’acquisizione di Arduino partirebbe con un’istruzione inviata dal computer tramite porta seriale in un istante *qualsiasi* dopo che lo script di Python è stato lanciato. In questo modo la misura sarebbe *asincrona* rispetto al processo sotto analisi, mentre invece è opportuno che l’acquisizione sia *sincrona* con, per esempio, l’inizio della fase di scarica. Il problema è concettualmente identico alla sincronizzazione della sweep dell’oscilloscopio tramite trigger. L’implementazione di un trigger “analogico”, cioè che monitora i livelli della grandezza digitalizzata, sarebbe molto poco efficiente dal punto di vi-

sta dei tempi di risposta, per cui la soluzione deve essere trovata altrove.

Il generatore di funzioni provvede un segnale pensato proprio per consentire la sincronizzazione. Questo segnale, che segue lo standard TTL [4], è presente su un’uscita con connettore BNC [5] posto sul frontale o sul retro dello strumento (dipende dai modelli) e riportata tramite cavo coassiale [6] su uno spinotto a banana di colore rosso (lo spinotto nero è collegato alla massa dello strumento e da qui alla terra dell’impianto elettrico). In particolare, come potrete facilmente verificare, questo segnale TTL è sincrono (stessa frequenza e fase coerente) con la forma d’onda prodotta dal generatore qualsiasi essa sia (quadra, sinusoidale, triangolare). Esso può facilmente essere “letto” da Arduino attraverso una delle porte digitali di cui è dotato (nell’esperimento si usa quella collegata con il pin 5) che, come dettagliato nello sketch riportato in Appendice, viene continuamente interrogata. Questo consente di far partire l’acquisizione in un istante che è approssimativamente coincidente con l’inizio della fase di scarica; ci sono ovviamente dei ritardi che però non mi sono mai deciso a misurare e che dovrebbero essere trascurabili per gli scopi dell’esperimento.

C’è poi un altro ottimo motivo per usare un’acquisizione sincrona. Come sapete, una delle principali limitazioni di Arduino nella sua versione Uno è la ridottissima capacità della memoria che registra i dati acquisiti. Nelle tipiche condizioni dei nostri esperimenti, in cui viene registrato il tempo e un canale digitalizzato, la memoria può contenere solo 256 “punti” sperimentali (cioè le misure sono fatte in 256 istanti successivi). Questo basso numero di punti potrebbe condurre a risultati poco soddisfacenti almeno dal punto di vista estetico. Inoltre la durata nominale dell’acquisizione, che è $\Delta t_{\text{acq}} = 256 \times \Delta t_{\text{samp}}$, potrebbe essere inadeguata per ricostruire le fasi di carica e scarica. Il problema può essere elegantemente aggirato sfruttando la periodicità dei processi considerati:

infatti è possibile ripetere N volte la stessa misura ($N = 8$ di default) in modo da ottenere record che contengono $N \times 256 = 8 \times 256 = 2048$ punti, che cioè coprono una durata complessiva nominale $\Delta t_{tot} = N \times \Delta t_{acq} = 8 \times \Delta t_{acq}$; chiaramente ogni singola acquisizione che va a costruire il record avviene in cicli successivi dell'onda quadra prodotta dal generatore, ma, grazie alla periodicità dei processi, è sufficiente che i dati acquisiti vengano correttamente "incollati" in sequenza per poter ricostruire correttamente gli andamenti. La procedura di "incollaggio" è effettuata dallo script di Python che controlla l'esperimento e si basa sul fatto che lo sketch di Arduino pone come istante iniziale dell'acquisizione n -esima l'istante finale dell'acquisizione $(n - 1)$ -esima. Naturalmente anche in questa valutazione dei tempi esistono delle incertezze, che però sono attese produrre effetti trascurabili (tutto potrebbe essere verificato sperimentalmente, ma non l'ho fatto).

Nonostante la configurazione di misura sia adeguatamente progettata, l'esito dell'esperimento richiede attenzione nel dimensionare i valori di R e C e nello scegliere i parametri di campionamento. In particolare:

- la durata complessiva del record Δt_{tot} , determinata da Δt_{samp} , deve permettere di ricostruire adeguatamente carica e scarica del condensatore, determinata approssimativamente da $\tau = RC$. In altre parole, è necessario che $V(t)$ sia campionata per una durata e con un numero di punti sufficienti per fare un buon best-fit. Tenete conto che in un best-fit secondo funzioni esponenziali le regioni asintotiche sono poco significative per la determinazione di τ , però dal punto di vista estetico è preferibile che almeno una porzione di asintoto sia presente.
- La frequenza f del generatore di funzioni, ovvero il periodo $T = 1/f$ dell'onda quadra usata nella carica/scarica del condensatore, deve essere adeguata agli altri parametri temporali. Infatti ogni singolo processo di carica o scarica può avvenire al massimo in un tempo pari a $T/2$.

III. ESEMPIO DI MISURE E BEST-FIT

Questa sezione riporta degli esempi di misure eseguiti per una determinata scelta di R e C (non ve la scrivo, perfidamente) e varie scelte dei parametri di lavoro Δt_{samp} e f .

La Fig. 2 mostra due record acquisiti per due distinte frequenze del generatore di funzioni, come indicato in legenda, e lo stesso valore dell'intervallo di campionamento nominale, $\Delta t_{sam} = 100 \mu s$. Un'analisi degli intervalli di campionamento effettivamente misurati da Arduino, eseguita sull'intero record, porta a $\Delta t_{sam} = (113 \pm 2) \mu s$, un valore maggiore rispetto a quello nominale impostato nello script di Python; la discrepanza è attesa e può essere interpretata tenendo conto del tempo necessario per completare la digitalizzazione che, come ben sapete, non può essere istantanea. Osservate come la deviazione standard

sperimentale sia minore dell'incertezza convenzionale pari a $\pm 4 \mu s$, a conferma che l'operazione di incollaggio di cui abbiamo trattato in precedenza non introduce errori apprezzabili sull'intero record [7].

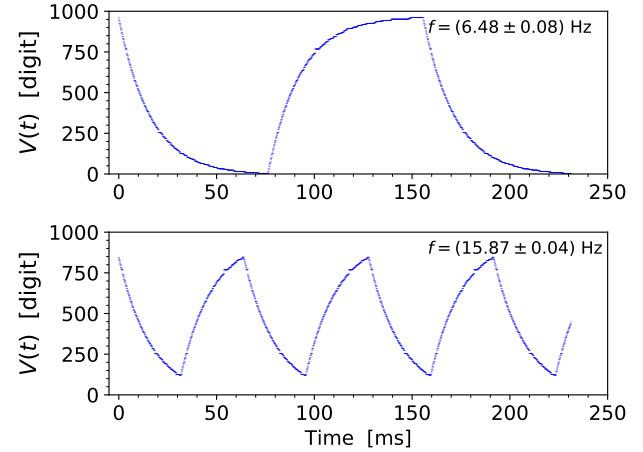


Figura 2. Grafici dei record acquisiti con la stessa scelta di R , C e Δt_{sam} e due diverse frequenze f del generatore di funzioni, come in legenda. Le barre di errore, determinate in modo convenzionale come specificato nel testo, sono presenti, ma scarsamente visibili a causa delle loro piccole dimensioni rispetto alla scala dei grafici.

Prima di procedere con l'analisi, sottolineiamo qualche aspetto rilevante:

- la frequenza del generatore di funzioni determina il numero di cicli di carica e scarica registrati nel record: è evidente che il record acquisito a frequenza più bassa consente una ricostruzione più fedele (con un numero maggiore di punti sperimentali) dei processi considerati.
- Si osserva chiaramente come all'aumentare della frequenza il segnale registrato tenda ad assumere una "forma" diversa e si attenui in ampiezza. La spiegazione è ovvia: il tempo $T/2$ a disposizione per carica e scarica diminuisce e il condensatore non fa in tempo a caricarsi e a scaricarsi completamente (cioè fino agli asintoti). Questo è il contenuto "scientifico" dell'ultima parte dell'esercitazione, che discuteremo nel seguito.
- Si osserva altrettanto chiaramente una particolarità di funzionamento decisamente fastidiosa, e non ben documentata in letteratura, del digitalizzatore di Arduino, almeno nella versione attualmente usata in laboratorio. Sono infatti visibili dei "buchi" nei record, in genere corrispondenti a valori digitalizzati prossimi a potenze di 2, per esempio $2^8 = 256$ digit, $2^9 = 512$ digit e soprattutto la loro somma, 768 digit. Tenteremo in seguito un'analisi un po' più approfondita del fenomeno; per il momento ci limitiamo a sottolineare che questo effetto,

che ha una natura sistematica, purtroppo limitata solo ad alcuni valori dell'intero range di misura, può rendere sottostimato l'errore di ± 1 digit da noi convenzionalmente impiegato.

La Fig. 3 riporta ancora i dati acquisiti per $f = (6.48 \pm 0.08)$ Hz con sovrapposti i risultati di due distinti best-fit eseguiti su scarica e carica. Le funzioni modello, scritte usando x e y per le variabili indipendenti e dipendenti, sono

$$y = A \exp(-x/\tau) + B \quad \text{scarica} \quad (1)$$

$$y = A \left(1 - \exp\left(-\frac{x-t'}{\tau}\right) \right) \quad \text{carica} ; \quad (2)$$

il pannello inferiore riporta il grafico dei *residui normalizzati* (cioè divisi per l'incertezza associata ai dati). I risultati dei best-fit, entrambi eseguiti con l'opzione `absolute_sigma = True` in ossequio all'origine presunta statistica degli errori convenzionali impiegati, sono riportati in Tab. I.

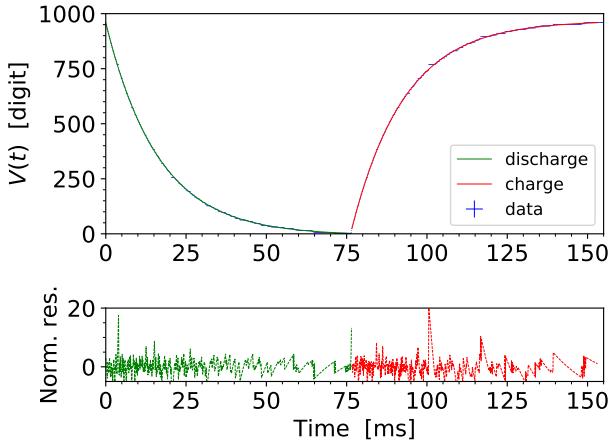


Figura 3. Grafico dei dati acquisiti per $f = (6.48 \pm 0.08)$ Hz, già mostrati nel pannello superiore di Fig. 2, con sovrapposti i best-fit delle fasi di scarica e carica realizzati come discusso nel testo. Il pannello inferiore riporta il grafico dei residui normalizzati dei best-fit.

Qualche osservazione su best-fit e risultati della misura:

- poichè la fisica dei processi è invariante rispetto all'unità di misura della d.d.p. campionata (il parametro maggiormente rilevante nel best-fit è τ , che per modello non dipende dall'unità di misura della d.d.p.), non c'è alcun bisogno di calibrare Arduino e di convertire i digit in V, con il vantaggio di poter trascurare gli errori di calibrazione. Tuttavia, per tenere conto dei possibili offset introdotti dal digitalizzatore, la funzione modello della fase di scarica comprende un parametro libero di offset, denominato B nell'equazione. Per come è scritta la funzione modello, questo parametro non può e

Tabella I. Risultati dei best-fit mostrati in Fig. 3; le ultime tre colonne riportano le covarianze normalizzate tra i vari parametri.

Scarica						
$\chi^2/ndof$	A [digit]	τ [μs]	B [digit]	$c_{A,\tau}$	$c_{A,B}$	$c_{\tau,B}$
4295/675	965.7 ± 0.2	16371 ± 6	-7.28 ± 0.08	-0.39	-0.07	-0.78
Carica						
$\chi^2/ndof$	A [digit]	τ [μs]	t' [ms]	$c_{A,\tau}$	$c_{A,t'}$	$c_{\tau,t'}$
6538/675	968.0 ± 0.3	16440 ± 7	76.20 ± 0.08	0.78	-0.79	-0.72

non deve essere inserito esplicitamente nell'equazione della carica, essendo di fatto compreso nel termine t' [ovvero $\exp(t'/\tau)$].

- Dato che anche la variabile indipendente è affetta da incertezza, per i best-fit, ovviamente realizzati minimizzando il “chi-quadro”, è stato impiegato il metodo dell'*errore efficace* a voi ben noto. Vista la presenza di funzioni esponenziali, l'uso di altri metodi più o meno abboracciati, tipo routine di minimizzazione ODR, *non è consentito*.
- I valori dei χ^2 ottenuti, che sono quelli che sono, risultano alti come possibile conseguenza della sottostima degli effetti sistematici di Arduino citati prima.
- I tempi caratteristici di carica e scarica non sono compatibili tra loro: questo può essere dovuto al comportamento elettronico del generatore di funzioni, la cui resistenza interna può effettivamente essere diversa (sperabilmente nell'ambito della tolleranza dichiarata dal costruttore) nelle due fasi. I tempi caratteristici sono invece ampiamente in accordo con i valori attesi, determinati attraverso misura di R , conoscenza di r_G e della capacità C , anch'essa nota con una tolleranza del 10%.
- I valori del parametro A , che ovviamente riflettono il valore massimo assunto da $V(t)$ nei processi, sono prossimi al massimo digitalizzabile da Arduino, cioè 1023 digit. Questo indica che l'ampiezza dell'onda quadra prodotta dal generatore è stata scelta in modo opportuno e che, come anche ben visibile a occhio, il condensatore ha avuto tempo sufficiente per approssimare le condizioni asintotiche di carica e scarica.
- I valori dei parametri B e t' (fasi rispettivamente di scarica e carica) sono in linea con le aspettative. L'offset B è infatti compatibile con le specifiche del digitalizzatore di Arduino e l'istante t' con l'istante di inizio della fase di carica (cioè $t' \simeq T/2$). Notate

che il termine di offset B potrebbe anche includere un contributo dovuto alla d.d.p. che l'onda quadra assume nel semiperiodo in cui si trova a livello basso (fase di scarica), che ho verificato essere pari a zero entro l'incertezza della misura con oscilloscopio.

- Le covarianze normalizzate, che hanno i segni attesi (per esempio, nella fase di carica A è correlato positivamente con τ , dato che al crescere di un parametro l'altro deve diminuire, e viceversa per le correlazioni di A e τ con t'), non mostrano “patologie” particolari. Per esempio, esse non si avvicinano pericolosamente a uno a indicare correlazioni (o anti-correlazioni, qui si intende il valore assoluto della covarianza normalizzata) pressoché complete tra parametri. Se questo fosse il caso, la funzione modello potrebbe essere stata scritta in forma sbagliata, per esempio con ridondanza di parametri. La bassa correlazione tra A e τ nella fase di scarica potrebbe essere dovuta al differente impatto numerico che il termine A , che moltiplica una funzione esponenziale, ha rispetto a B .

Come già preannunciato, l'esperimento costituisce un possibile metodo, piuttosto involuto e un po' complicato, per la misura indiretta della capacità C . Poiché $C = \tau/(R + r_G)$, l'incertezza della misura dipende da quella su τ ottenuta dal best-fit e da quella sulla somma $(R + r_G)$ (l'incertezza su R è data dalla misura, quella su r_G dalla tolleranza dichiarata nel manuale): i due contributi di incertezza relativa, che compaiono a numeratore e denominatore dell'espressione, possono essere sommati in quadratura tra loro essendo supposti indipendenti. Nel mio esempio ho scelto $R \gg r_G$ e misurato R con il multimetro digitale ordinario ottenendo un errore relativo di circa 0.9%, che domina ampiamente sull'errore relativo di τ (vale circa 4×10^{-4} , ovvero 0.04%, ovvero ancora circa 400 ppm) e sulla presenza di r_G (e quindi della sua incertezza). Quindi la mia misura di C è risultata affetta da questo stesso errore relativo, e ovviamente in ampio accordo con le aspettative, data la grande tolleranza (10%) con cui il costruttore ha dichiarato la capacità.

Ho comunque anche ripetuto la misura di R usando il multimetro digitale Fluke 8808A, basato su microvoltmetro, disponibile in laboratorio. In questo caso ho determinato R con un'incertezza relativa di circa 3×10^{-4} (0.03%, ovvero circa 300 ppm), che comporta di dover tenere conto della presenza di r_G e della sua tolleranza. L'incertezza relativa sulla somma $(R + r_G)$ diventa di circa 5×10^{-4} (0.05%, ovvero circa 500 ppm) che è paragonabile a quella su τ . Questo consente di ottenere una misura indiretta di C con un'incertezza relativa di circa 7×10^{-4} (0.07%, ovvero circa 700 ppm), che non è niente male!

A. Problemi di Arduino e outliers

Come già affermato e come sempre dovrebbe essere, il χ^2 ottenuto dal best-fit è quello che è, dato che il suo valore dipende dalla stima delle incertezze di misura e queste non possono essere determinate con affidabilità in mancanza di modelli accurati del funzionamento del digitalizzatore, che potrebbero essere per esempio impiegati in simulazioni numeriche, o di misure dirette degli errori, che richiederebbero almeno la presenza di sorgenti calibrate di d.d.p., oltre alla messa a punto di opportune, e presumibilmente complesse, procedure.

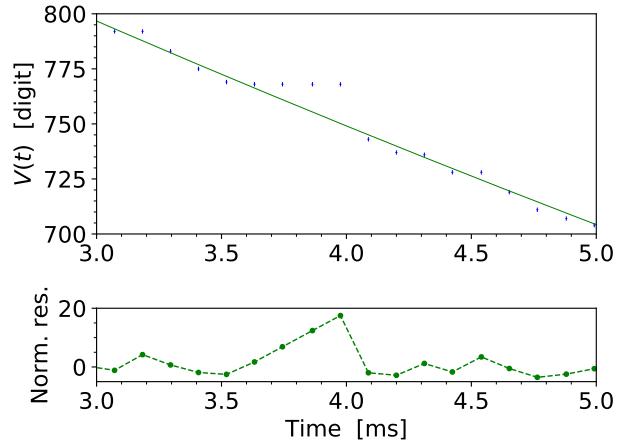


Figura 4. Espansione dei grafici mostrati in Fig. 3, fase di scarica, attorno a valori digitalizzati di 768 digit.

La Fig. 4 mostra un'espansione del grafico di Fig. 3 in cui viene evidenziato il comportamento nella fase di scarica quando il valore digitalizzato si avvicina a 768 digit: si vede chiaramente che il digitalizzatore, nelle condizioni in cui lo impieghiamo, “non riesce a seguire” l'andamento previsto dalla scarica, rappresentato dalla curva continua verde ottenuta dal best-fit. Di fatto è come se esso “si incantasse”: la lettura rimane bloccata a 768 digit per un certo numero di punti (4 o 5, nel caso considerato) e poi riparte regolarmente. Di conseguenza i residui normalizzati, che nel caso di errori statistici distribuiti in modo normale dovrebbero essere prevalentemente compresi tra -1 e +1, esplodono con un andamento di tipo esponenziale (la differenza tra il valore costante registrato da Arduino e la funzione modello). Il fenomeno, non ben documentato in letteratura, potrebbe essere dovuto all'architettura interna di Arduino dove vengono usati diversi digitalizzatori in cascata tra loro, la cui attivazione avviene in corrispondenza a valori digitalizzati prossimi a potenze di 2, o loro somme (in effetti situazioni simili si riscontrano anche attorno ad altri valori digitalizzati, per esempio 256 e 512 digit). Come ricordate, un digitalizzatore ha bisogno di un circuito sample-and-hold che mantenga costante la d.d.p. da misurare durante l'intero processo di digitalizzazione. Potrebbe essere che, attorno

a questi valori speciali, Arduino sia troppo impegnato a commutare tra i digitalizzatori in cascata e si dimentichi di controllare opportunamente il sample-and-hold.

Osservate che un comportamento “seghettato” si ritrova anche altrove nei grafici, per esempio nella fase finale di scarica evidenziata in Fig. 5: in questo caso il digitalizzatore potrebbe stentare a seguire l’andamento temporale del segnale anche perché questo si avvicina agli estremi del range che può essere digitalizzato, comportamento che si ritrova anche in dispositivi più sofisticati di Arduino e (solo in parte!) interpretabile con le fluttuazioni della lettura: per intenderci, assumendo un’incertezza ± 1 digit, nel caso estremo una d.d.p. di valore 1 digit potrebbe dare luogo a letture pari a 0, 1, 2 digit. Notate comunque che, in questi casi, normalmente i residui non esplodono come osservato in precedenza, pur mantenendone lo stesso andamento.

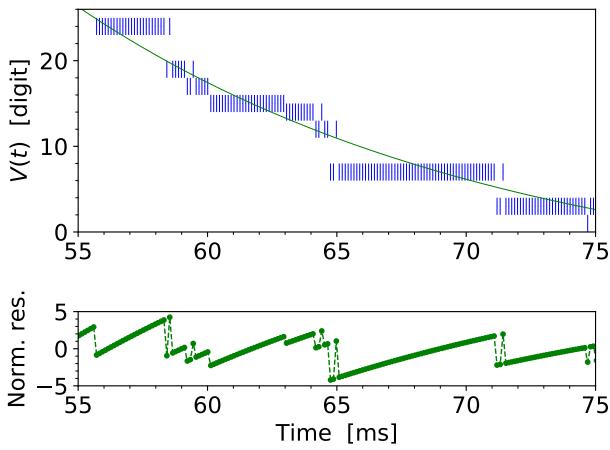


Figura 5. Espansione dei grafici mostrati in Fig. 3 attorno al termine della fase di scarica.

Vista la sintomatologia, è evidente che una terapia efficace per questi problemi, in particolare il primo, non può essere trovata facilmente. Magari si potrebbe aumentare arbitrariamente l’incertezza per i campionamenti prossimi ai valori critici di digitalizzazione. Per mantenersi nel territorio dell’arbitrarietà, discutiamo qui nel seguito un metodo che può essere qualche volta impiegato, a patto di essere *sempre dichiarato*, per ragioni che sono fondamentalmente di tipo estetico, cioè per soddisfare il vostro bisogno estetico di un χ^2 “non troppo grande”. Il metodo si basa sull’*identificazione e soppressione degli effetti dei cosiddetti outliers*.

Gli outliers sono in pratica quei punti sperimentali che deviano significativamente dalle aspettative. Identificare gli outliers e sopprimerne gli effetti nel best-fit è una pratica comune nel caso di valori digitalzzati decisamente fuori scala che possono essere attribuiti a motivi statistici. Per esempio, l’arrivo di radiazione cosmica ionizzante all’interno del digitalizzatore può dare luogo a “spikes” di d.d.p. e lo stesso può verificarsi, per induzione magnetica

o elettrica, quando nel resto del circuito ci sono brusche variazioni di d.d.p. o correnti: i più attenti tra voi noteranno che, per limitare quest’ultimo problema, i dati corrispondenti agli istanti in cui l’onda quadra cambia il proprio livello vengono scartati nell’acquisizione (si veda lo sketch in Appendice).

Qui, in modo arbitrario e per motivi estetici, applichiamo il metodo degli outliers a deviazioni dalle previsioni che hanno un’origine non puramente statistica, dato che i valori critici della digitalizzazione sono determinati. L’occasione è utile anche per progettare e creare il codice necessario, che viene ovviamente lasciato per esercizio.

La Fig. 6 riporta gli stessi dati di Fig. 3 per la fase di scarica, la sola considerata qui per brevità: notate che in questo caso è stata adottata la rappresentazione semi-logaritmica per $V(t)$ [8]. In questo esempio sono stati arbitrariamente identificati come outliers quei dati sperimentali che distavano più di tre barre di errore rispetto all’aspettativa, cioè al valore previsto dal best-fit [9]. Come strettamente necessario (*obbligatorio*), gli outliers *non* sono rimossi dal grafico, dove rimangono essendo evidenziati in modo chiaro (markers viola in figura). Infatti è generalmente vietato manipolare i dati acquisiti cancellandone alcuni! La Tab. II riassume i risultati del best-fit eseguito senza considerare gli outliers e sempre con l’opzione `absolute_sigma = True`: dal confronto dei gradi di libertà (*ndof* nelle tabelle) con i risultati di Tab. I si vede come oltre 100 dati sperimentali siano stati classificati come outliers e non considerati nel best-fit. Si osserva inoltre un’ovvia diminuzione del χ^2 , con un $\chi^2_{rid} = \chi^2/ndof$ che si avvicina ai valori ordinari per best-fit su dati con errori di origine statistica. Non tutti i valori dei parametri sono compatibili con quelli trovati nel best-fit ordinario, ma le differenze non conducono a variazioni significative nella misura indiretta di C .

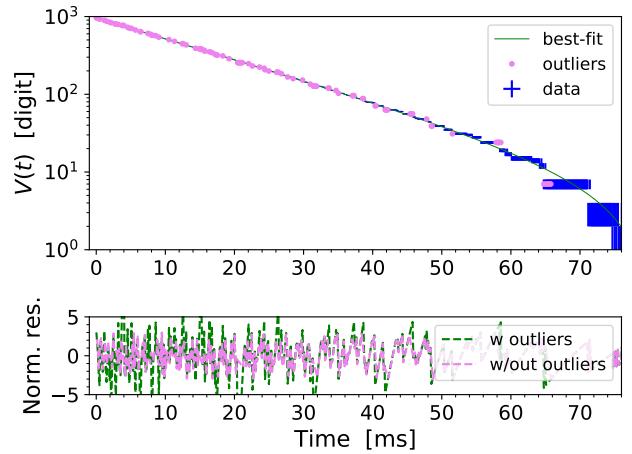


Figura 6. Stessi dati mostrati in Fig. 3, fase di scarica, con indicazione degli outliers individuati come discusso nel testo. Il pannello inferiore riporta il grafico dei residui normalizzati dei best-fit considerando, o meno, gli outliers tra i dati.

Tabella II. Risultati del best-fit mostrato in Fig. 6, eseguito sugli stessi dati di Fig. 3, ma escludendo gli outliers identificati come discusso nel testo; le ultime tre colonne riportano le covarianze normalizzate tra i vari parametri.

Scarica						
$\chi^2/ndof$	A [digit]	τ [μs]	B [digit]	$c_{A,\tau}$	$c_{A,B}$	$c_{\tau,B}$
1167/565	965.3 ± 0.2	16393 ± 7	-7.41 ± 0.01	-0.40	-0.06	-0.78

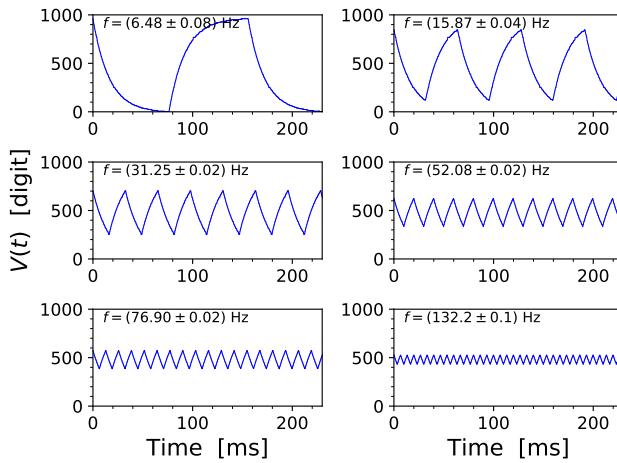


Figura 7. Segnale $V(t)$ registrato per diversi valori di f del generatore di funzioni, come in legenda; tutti gli altri parametri sono invariati rispetto a quanto illustrato in precedenza nel testo.

IV. FORME D'ONDA IN FUNZIONE DELLA FREQUENZA

Un altro scopo “scientifico” dell’esercitazione pratica riguarda l’acquisizione del segnale $V(t)$ eseguita variando

su un range piuttosto ampio la frequenza f del generatore (tutto il resto viene lasciato inalterato e si lascia per utile esercizio la costruzione di analoghi record con l’impiego di forme d’onda in ingresso sinusoidali e triangolari).

La Fig. 7 mostra un esempio di dati acquisiti su un range di frequenze che spazia per oltre una decade (potete usare un range ancora più ampio!). I record acquisiti saranno poi analizzati con tecniche un po’ più avanzate (serie e trasformata di Fourier) più avanti nel corso. Per il momento limitiamoci a sottolineare ancora che all’aumentare della frequenza l’ampiezza dell’oscillazione registrata diminuisce e che la forma cambia, passando da qualcosa di molto simile a un’onda quadra a una sorta di “pinna di squalo” e infine a qualcosa che somiglia molto a un’onda triangolare; interpreteremo poi la forma triangolare come “integrale nel tempo” dell’onda quadra in ingresso.

APPENDICE: SKETCH DI ARDUINO E SCRIPT DI PYTHON

Per l’esercitazione pratica si usa una combinazione di sketch e script, di nome `synclong2016` (con estensione `.ino` e `.py` per sketch e script, rispettivamente) originalmente creata per altri scopi: non escludo di modificarli a breve. Essi sono disponibili nei computer di laboratorio (directory `Arduini`) e in rete nel sito di e-learning del corso.

La lettura delle poche righe di codice, assieme ai commenti (sgrammaticati) scritti qua e là, è raccomandata a tutti; la comprensione delle varie istruzioni è riservata alla frazione più smanettona e appassionata di coding, if any.

```
/*
Questo sketch serve per acquisire forme d'onda dipendenti dal tempo in modalita'
glued. Vengono eseguite in successione 8 acquisizioni con un ritardo variabile
rispetto al trigger. La porta analogica letta A0 e il trigger digitale e' sulla porta 5.
In ogni acquisizione vengono acquisiti 256 punti. Il files complessivo avra'
256x8=2048 righe e due colonne (rispettivamente t in us e valore digitalizzato).
*/
// Blocco definizioni
const unsigned int analogPin=0; // Definisce la porta A0 per la lettura
const unsigned int sincPin = 5; //pin 5 ingresso digitale per la sincronizzazione con il generatore
int i; // Definisce la variabile intera i (contatore)
int delays; // Definisce la variabile intera delays
int V[256]; // Definisce l'array intero V
long t[256]; // Definisce l'array t
unsigned long StartTime; // Definisce il valore StartTime
```

```

unsigned long delayus; // Definisce variabile per acquisizione multipla
unsigned long delayms; // Definisce variabile ausiliaria tempo totale acq
int start=0; // Definisce il valore start (usato come flag)
int sinc; // Variabile di sincronizzazione
int j; // Variabile di loop multiacquisizione

// Istruzioni di inizializzazione
void setup()
{
    Serial.begin(19200); // Inizializza la porta seriale a 19200 baud
    Serial.flush(); // Pulisce il buffer della porta seriale
    pinMode(sincPin, INPUT); //pin sincPin configurato come ingresso digitale
    //analogReference(INTERNAL); // Sceglie il riferimento V_ref = 1.1 V (nominali)
    bitClear(ADCSRA,ADPS0); // Istruzioni necessarie per velocizzare
    bitClear(ADCSRA,ADPS2); // il rate di acquisizione analogica
}

// Istruzioni del programma
void loop()
{
    delayus=0; // Valori iniziali variabili di ritardo
    delayms=0;

    if (Serial.available() >0) // Controlla se il buffer seriale ha qualcosa
    {
        delays = (Serial.read()-'0')*100; // Legge il byte e lo interpreta come ritardo
        Serial.flush(); // Svuota la seriale
    start=1; // Pone il flag start a uno
    }

    if(!start) return // Se il flag e' start=0 non esegue le operazioni qui di seguito
                    // altrimenti le fa partire (quindi aspetta di ricevere l'istruzione
                    // di partenza)
    delay(1000); // Aspetta 1000 ms per evitare casini
    for (j=0;j<8;j++)
    {
//    delay(1000); // Aspetta 1000 ms per evitare casini
        sinc = digitalRead(sincPin);//legge sincPin
        while (sinc==HIGH) // ciclo di attesa iniziale per sincronizzazione, attende che sincPin vada basso
            {sinc = digitalRead(sincPin);} //legge sincPin
        while (sinc==LOW) // ciclo di attesa iniziale per sincronizzazione, attende che sincPin vada alto
            {sinc = digitalRead(sincPin);} //legge sincPin
        StartTime=micros(); // Misura il tempo iniziale con l'orologio interno
        if (j==0)
        {
            for(i=0;i<2;i++) // Loop di due misure a vuoto (da fare solo all'inizio)
            {
                V[i]=analogRead(analogPin);
            }
        }
        if (j>0)
        {
            delayMicroseconds(delayus+delays);
            delay(delayms);
        }
    for(i=0;i<256;i++) // Loop di misura
    {
        t[i]=micros()-StartTime; // Legge il timestamp e lo mette in array t
    }
}
}

```

```

V[i]=analogRead(analogPin); // Legge analogPin e lo mette in array V
delayMicroseconds(delays); // Aspetta tot us
}
delayms=floor(t[255]/1000);
delayus=t[255]-delayms*1000;
for(i=0;i<256;i++) // Loop per la scrittura su porta seriale
{
    Serial.print(t[i]); // Scrive t[i]
    Serial.print(" "); // Mette uno spazio
    Serial.println(V[i]); // Scrive V[i] e va a capo
}
delay(1000); // Aspetta 1000 ms per evitare casini
}

start=0; // Annulla il flag
Serial.flush(); // Pulisce il buffer della porta seriale (si sa mai)
}

```

```

# Questo script serve per interfacciarsi con Arduino nell'esperienza
# dell'acquisizione sincrona di segnali
# L'interfacciamento avviene attraverso:
# 1. scrittura di un carattere (byte) che esprime l'intervallo di campionamento
# 2. lettura dei dati disponibili su porta seriale

import serial # libreria per gestione porta seriale (USB)
import time # libreria per temporizzazione

Directory='../../dati_arduino/' # nome directory dati
                                # << DA CAMBIARE SECONDO NECESSITA'
FileName='datilunghi.txt' # nome file << DA CAMBIARE SECONDO NECESSITA'

ard=serial.Serial('/dev/ttyACM0',19200) # apre porta seriale (occhio alla sintassi, dipende
                                         # dal sistema operativo!)
time.sleep(2) # aspetta due secondi per evitare casini

ard.write(b'1') # scrive il carattere per l'intervallo di campionamento
                 # in unita' di 100 us << DA CAMBIARE A SECONDA DEI GUSTI
                 # l'istruzione b indica che e' un byte (carattere ASCII)

time.sleep(1) # aspetta un secondo per evitare casini

outputFile = open(FileName, "w" ) # apre file dati carica per scrittura

print ("start")

for j in range (0,8): # loop acquisizioni multiple (sono 8 di default)

# loop lettura dati da seriale (256 punti)
    for i in range (0,256):
        data = ard.readline().decode() # legge il dato e lo decodifica
        if data:
            outputFile.write(data) # scrive i blocchi di dati
    print ("Part ",j+1," done")

```

```

outputFile.close() # chiude il file dei dati di carica

ard.close() # chiude la comunicazione seriale con Arduino

print('end') # scrive sulla console che ha finito

```

- [1] Vedremo nel seguito che esistono altri modi, spesso più efficaci, per misurare la capacità, per esempio attraverso l'analisi della risposta di filtri RC , circuiti RLC risonanti, per confronto con capacità note usando ponti di de Sauty.
- [2] L'affermazione è valida solo in prima approssimazione. Infatti, per motivi sia tecnici che fondamentali, è impossibile che la d.d.p. cambi di livello istantaneamente. Per il generatore di funzioni impiegato in laboratorio il manuale indica un tempo di salita o discesa dell'onda quadra inferiore a 100 ns, un valore molto minore rispetto alle scale temporali di interesse per l'esperimento. Quindi l'approssimazione è ampiamente giustificata.
- [3] Naturalmente la presenza di r_o può essere tenuta in conto nel modello; tuttavia, dato che questa resistenza si trova in parallelo al condensatore, la soluzione analitica del circuito diventa molto complicata e l'equazione modello poco maneggevole, motivo per cui rinunciamo volentieri a modellarne gli effetti.
- [4] Un segnale TTL (Transistor Transistor Logic) può assumere solo due livelli, "alto" o "basso" rispetto alla linea di riferimento, tipicamente la massa o terra. Approssimativamente (esistono diverse varianti per questo standard) il livello alto corrisponde a una d.d.p. di circa 5 V, il livello basso a una d.d.p. quasi nulla. Per completezza segnaliamo che lo stesso segnale può anche assumere valori diversi quando viene attivata, mediante estrazione, la manopola TTL/CMOS posta sul frontale del generatore di funzioni; CMOS è infatti la denominazione di un altro standard ampiamente adottato in elettronica.
- [5] Per ora si tratta di una sigla di cui più avanti conosceremo il significato.
- [6] Per ora si tratta di un aggettivo di cui più avanti conosceremo il significato!
- [7] L'incertezza convenzionale che attribuiamo alle misure temporali di Arduino, che è supportata dalla documentazione tecnica disponibile, non tiene conto dell'errore di calibrazione del clock. Questo errore non è noto, ma può essere desunto dall'indicazione della frequenza riportata sul "quarzo" impiegato da Arduino. L'indicazione è 16.000 (si intende che l'unità di misura è MHz). Possiamo quindi dedurre che l'errore relativo nella misura dei tempi dovuto alla calibrazione è di $1/16000 = 62.5$ ppm (vuol dire parti per milione!); in altre parole, nel tempo caratteristico di carica e scarica del circuito analizzato nel testo l'incertezza di calibrazione porta a un errore di circa un microsecondo, che riteniamo trascurabile per i nostri scopi. Riteniamo trascurabile anche l'eventuale effetto delle variazioni di temperatura, le quali sono in generale attese influire sulla frequenza di oscillazione del quarzo (si tratta alla fine di un oscillatore meccanico!).
- [8] La rappresentazione semi-logaritmica è quanto mai appropriata per i dati considerati in figura, i cui valori spaziano su diversi decadi. Come potete facilmente verificare, un andamento esponenziale decrescente risulta rettilineo in rappresentazione semi-logaritmica, con una pendenza negativa proporzionale a $1/\tau$; nel caso di figura l'andamento appare curvato a causa del termine di offset (B nella funzione modello) non nullo.
- [9] Poiché si tratta di una previsione, l'incertezza sul valore del best-fit dovrebbe essere correttamente calcolata considerando le covarianze, secondo il metodo che conoscete. Per semplicità e pigrizia, questo non è stato fatto. Notate tuttavia che l'incertezza utilizzata, come sempre nel calcolo dei residui normalizzati presentati in questa nota, è l'errore efficace determinato per propagazione dell'errore della misura dei tempi sulla funzione modello.

Filtri RC e Python

francesco.fuso@unipi.it

(Dated: version 10 - FF, 28 novembre 2019)

Questa nota riporta istruzioni e commenti sulla costruzione della curva di risposta e del diagramma di Bode di un filtro RC usando Python e fa riferimento all'esercitazione pratica di progettazione e costruzione di filtri RC "a un polo" (passa-basso o passa-alto, con un solo elemento resistivo e un solo elemento capacitivo).

I. INTRODUZIONE

I filtri sono in generale dei dispositivi che "lasciano passare" pressoché inalterati segnali periodici in un certo intervallo di frequenza e "modificano", cioè attenuano e sfasano, segnali periodici in un altro intervallo di frequenza. Nello specifico, un filtro passa-basso attenua e sfasa segnali di frequenza al di sopra di un certo valore, un filtro passa-alto attenua e sfasa quelli di frequenza al di sotto di un certo valore. Con un ragionamento naïf si potrebbero ipotizzare delle curve di risposta "a gradino", in cui il guadagno cambia in modo brusco a seconda che si sia al di sotto o al di sopra di una certa frequenza (o viceversa, a seconda del tipo di filtro).

Come sarà illustrato brevemente in questa nota, il comportamento dei filtri reali non segue questa descrizione, prevedendo una "regione di transizione" nella quale guadagno e sfasamento dipendono dalla frequenza, dando luogo a curve di risposta che sono "smussate" rispetto al gradino: ciò può essere facilmente modellato nel caso dei filtri RC, determinando la cosiddetta *funzione di trasferimento*, o spettro di risposta, del circuito $T(\omega)$, generalmente complessa.

Nell'ambito del metodo simbolico, cioè supponendo segnali sinusoidali in uscita e in ingresso ed usando la forma fasoriale per esprimere, si ha: $V_{\omega,out} = T(\omega)V_{\omega,in}$. Le ampiezze (ovvero i valori picco-picco, il comportamento è ovviamente lo stesso per ampiezze e picco-picco) sono determinate dai moduli dei fasori considerati: $|V_{\omega,out}| = |T(\omega)V_{\omega,in}| = A(\omega)|V_{\omega,in}|$, dove abbiamo introdotto la funzione (reale) $A(\omega) = |T(\omega)|$ che, essendo pari al rapporto tra ampiezza in uscita e ampiezza in ingresso, mostra proprio quanto vale il *guadagno* in ampiezza all'uscita rispetto all'ingresso. Per i filtri passivi, che non includono componenti in grado di amplificare, è sempre $A(\omega) \leq 1$, motivo per cui si usa qualche volta il termine *attenuazione* invece che guadagno.

Come "spunto di riflessione", nel senso che l'argomento non viene trattato nell'ambito di esercitazioni pratiche ordinarie, potete pensare a come potrebbe essere realizzato un *filtro passa-banda* (o un *filtro notch*, ovvero attenutabanda), cioè un dispositivo che fa passare, o attenua e sfasa, solo i segnali che si trovano all'interno di una certa banda di frequenze [1]. Siete invitati a farlo cercando di passare da un atteggiamento naïf a un ragionamento nel quale individuate i limiti tecnici delle configurazioni più semplici. In questo potrebbe farvi comodo riflettere sul

circuito in cascata integratore/derivatore protagonista di una precedente esercitazione pratica.

A. decibel

Molto spesso si ha a che fare con guadagni che hanno valori numerici molto piccoli o molto grandi. Un'unità di misura opportuna in tali situazioni è il *decibel* (dB). Il guadagno *in ampiezza* del segnale in dB è definito come

$$A(\omega) [\text{dB}] = 20 \log_{10} \left(\frac{V_{out}}{V_{in}} \right), \quad (1)$$

dove V_{out} e V_{in} rappresentano le *ampiezze* dei segnali in uscita e in ingresso.

Occorre precisare si dà una diversa definizione (con un fattore 10 al posto del fattore 20) nel caso in cui il guadagno si riferisca a *potenze* (o energie): queste sono infatti generalmente proporzionali al quadrato delle ampiezze corrispondenti, e la doppia definizione permette di avere coerenza numerica fra guadagni in ampiezza e in potenza. In altre parole, grazie alla doppia definizione un guadagno di -20 dB implica attenuazione di un fattore 10 sia in ampiezza che in potenza. Inoltre, e purtroppo in una forma non molto ben codificata, qualche volta si usa una scala logaritmica anche per indicare il valore "assoluto" di ampiezze o potenze. A questo scopo l'ampiezza del segnale viene divisa per un'unità di riferimento (in sostanza questa unità di riferimento prende il posto di V_{in} nell'Eq. 1 o nella sua analoga per la potenza): nel caso in cui l'unità di riferimento sia 1 mV (o 1 mW, per la potenza), l'unità di misura si scrive *in genere* dBm, ma sono purtroppo presenti anche altre convenzioni.

II. PASSA-BASSO E PASSA-ALTO

Un filtro passa-basso di quelli considerati qui è composto dalla serie di un resistore e di un condensatore. L'ingresso è collegato al generatore di forme d'onda, l'uscita misura la d.d.p. ai capi del condensatore.

La funzione di trasferimento complessa per il circuito considerato (stiamo trascurando la resistenza interna del generatore e gli effetti di quella dello strumento di misura, come disucssso anche nella seguente sezione) è del tipo:

$$T(\omega) = \frac{1}{1 + j\omega RC} = \frac{1}{1 + j\omega/\omega_T}, \quad (2)$$

dove abbiamo definito la frequenza angolare di taglio $\omega_T = 1/(RC)$. Ricordando il legame tra frequenza angolare e frequenza, $\omega = 2\pi f$, e definendo la *frequenza di taglio* $f_T = \omega_T/(2\pi) = 1/(2\pi RC)$, si ha anche

$$T(f) = \frac{1}{1 + jf/f_T}. \quad (3)$$

Lo sfasamento $\Delta\phi$ tra i fasori $V_{\omega,out}$ e $V_{\omega,in}$ è legato al rapporto tra parte immaginaria e parte reale della funzione di trasferimento, cioè

$$\tan(\Delta\phi) = \frac{\text{Im}\{T(f)\}}{\text{Re}\{T(f)\}} = -\frac{f}{f_T}, \quad (4)$$

Si ha $\Delta\phi \rightarrow 0$ per $f \rightarrow 0$, $\Delta\phi \rightarrow -\pi/2$ per $f \rightarrow \infty$; inoltre è $\Delta\phi_T = -\pi/4$ per $f = f_T$. Notate che i segni negativi che compaiono negli sfasamenti richiedono attenzione per essere determinati dal punto di vista sperimentale, e spesso è sufficiente, e generalmente più semplice, misurarli in valore assoluto.

Inoltre per un filtro passa-basso è

$$A(f) = |T(f)| = \frac{1}{\sqrt{1 + (f/f_T)^2}}. \quad (5)$$

Si vede facilmente che $A(f) \rightarrow 1$ per $f \rightarrow 0$, $A(f) \rightarrow 0$ per $f \rightarrow \infty$; inoltre è $A_T = 1/\sqrt{2}$ per $f = f_T$. Spesso si definisce un'ulteriore frequenza caratteristica dei filtri, $f_{1/2}$, che equivale alla frequenza a cui il guadagno, o attenuazione, vale $A(f_{1/2}) = 1/2$. Si vede facilmente che $f_{1/2} = \sqrt{3}f_T$ (e inoltre a tale frequenza si ha $\Delta\phi_{1/2} = -\pi/3$). Ragionando in termini di dB, alla frequenza di taglio il guadagno vale $20 \log_{10}(1/\sqrt{2}) \approx -3$ dB, e alla frequenza $f_{1/2}$ esso vale $20 \log_{10}(1/2) \approx -6$ dB.

Un filtro passa-alto di quelli qui considerati è fatto ancora dalla serie di un condensatore e di un resistore, ma stavolta l'uscita è presa ai capi del resistore. La funzione di trasferimento è, in questo caso

$$T(f) = \frac{1}{1 - jf_T/f}, \quad (6)$$

a cui corrispondono la funzione di risposta per il guadagno

$$A(f) = |T(f)| = \frac{1}{\sqrt{1 + (f_T/f)^2}}, \quad (7)$$

e lo sfasamento

$$\tan(\Delta\phi) = \frac{\text{Im}\{T(f)\}}{\text{Re}\{T(f)\}} = \frac{f_T}{f}. \quad (8)$$

Si vede facilmente che $A(f) \rightarrow 1$ per $f \rightarrow \infty$, $A(f) \rightarrow 0$ per $f \rightarrow 0$; inoltre è $A_T = 1/\sqrt{2}$ per $f = f_T$ e $A_{1/2} = 1/2$ per $f_{1/2} = f_T/\sqrt{3}$ (e inoltre a tale frequenza si ha $\Delta\phi_{1/2} = \pi/3$).

III. MISURE SUL FILTRO PASSA-BASSO

Ho costruito e fatto misure [2] su un filtro passa-basso realizzato con $R = (3.27 \pm 0.03)$ kohm (misurata con multimetro digitale) e $C = 0.1 \mu\text{F}$ con tolleranza 10%. Per questo filtro ci si aspetta $f_{T,att} = (0.49 \pm 0.05)$ kHz, essendo l'incertezza dovuta principalmente alla tolleranza sulla capacità. Notate che l'uso di un valore di resistenza ben più alto della resistenza interna del generatore ($r_G = 50$ ohm, nominale) ci permette di considerare *approssimativamente* trascurabile quest'ultima. Infatti l'*impedenza* del circuito vale, in modulo, $|Z_{tot}| = \sqrt{R^2 + 1/(\omega C)^2}$, che, per la scelta fatta, è sempre ben maggiore di r_G a qualsiasi frequenza esplorata nell'esperimento [3]. Inoltre in queste stesse condizioni l'*impedenza* del condensatore, pari in modulo a $1/(\omega C)$, è sempre minore della *impedenza interna* dell'oscilloscopio [4], per cui la presenza dello strumento di misura produce effetti trascurabili rispetto all'accuratezza tipica delle misure (ricordate il 3% tipico di incertezza di calibrazione).

In primo luogo ho determinato sperimentalmente la frequenza di taglio f_T : a questo scopo ho misurato l'ampiezza picco-picco del segnale in ingresso $V_{in} = (9.9 \pm 0.4)$ V, dove l'incertezza tiene conto dell'errore di calibrazione dato dal costruttore e della difficoltà di posizionare correttamente i cursori dell'oscilloscopio sui picchi, cioè dello spessore della traccia. Quindi ho cercato al variare della frequenza (girando lentamente la manopola del generatore di funzioni) il valore per cui l'ampiezza di V_{out} si riduceva di un fattore $1/\sqrt{2}$ rispetto a questa. Ho ottenuto $f_T = (480 \pm 10)$ Hz, valore in accordo con le attese. L'incertezza qui è dovuta non solo alla precisione con cui si determina la frequenza (il frequenzimetro integrato nell'oscilloscopio ha una precisione nominale dello 0.05% nel range considerato, quella del frequenzimetro integrato nel generatore di forme d'onda, debitamente terminalizzato, è 20 ppm, oltre all'incertezza di lettura pari a ± 1 digit) o alla stabilità del generatore (che fa "ballare" qualche cifra significativa sul display dei frequenzimetri), ma soprattutto all'errore che si compie nell'individuare la frequenza a cui si verifica la condizione prescritta. Questa situazione è tipica di quando si eseguono misure "indirette", o "condizionate": infatti basta sfiorare la manopola del generatore di forme d'onda per vedere sensibili cambiamenti di frequenza, pur mantenendosi pressoché inalterata la condizione richiesta sull'ampiezza. Ho ripetuto varie volte la misura e stimato di conseguenza l'incertezza dallo scarto dei risultati. Ho anche misurato lo sfasamento a questa frequenza, ottenendo $\Delta\phi_T = -(0.26 \pm 0.03)$ π rad, dove l'incertezza è dovuta principalmente alla difficoltà di individuare l'istante in cui le tracce passano per il livello di zero (le tracce sono spesse).

Inoltre ho individuato la frequenza $f_{1/2}$ alla quale l'ampiezza in uscita è la metà di quella in ingresso, trovando $f_{1/2} = (876 \pm 12)$ Hz [sfasamento corrispondente $\Delta\phi_{1/2} = -(0.32 \pm 0.03)\pi$ rad]. I valori di f_T e $f_{1/2}$ sono in accordo con le attese basate sui valori nominali di progetto. Inoltre gli sfasamenti misurati sono anche in accordo

con le aspettative (rispettivamente $-\pi/4$ e $-\pi/3$).

Ho quindi misurato le ampiezze picco-picco $V_{in,j}$ e $V_{out,j}$ del segnale in uscita dal filtro a diverse frequenze f_j e ho quindi fatto determinare a Python l'attenuazione $A_j = V_{out,j}/V_{in,j}$. Avendo in mente di dover spazzare un range piuttosto ampio e di dover produrre un grafico in carta bilogaritmica, ho deciso di acquisire i dati a frequenze via via crescenti, l'una circa pari al doppio della precedente (cioè ho variato le frequenze di circa un'ottava alla volta), in modo da produrre un grafico in cui le frequenze fossero pressoché equispaziate. I dati sono mostrati in Fig. 1: a un grafico che, come quello mostrato, riporta la frequenza (o, come vedremo, energia, o lunghezza d'onda) nell'asse orizzontale si dà spesso il nome gergale di *spettro*, e quindi il grafico può essere definito “lo spettro del guadagno” o “della risposta” del filtro costruito. Per le incertezze ΔA_j ho propagato l'incertezza delle misure di ampiezza, in cui ho tenuto conto sia della calibrazione (dato che ho usato diverse scale di portata su tutti i due canali dell'oscilloscopio [5]) che dell'errore di lettura (cursori e spessore della traccia). Per la misura delle frequenze ho considerato un'incertezza pressoché trascurabile (nettamente inferiore all'1%), stimata osservando la cifra ballerina dei display del frequenzimetro del generatore di forme d'onda.

Nell'eseguire l'esperimento ho notato un aspetto fastidioso dal punto di vista sperimentale, legato alla progressiva riduzione del valore di ampiezza dell'onda prodotta dal generatore (a prescindere dal carico, è un problema di mancata stabilizzazione in temperatura) nel corso delle misure. Per minimizzarne gli effetti ho cercato di eseguire le misure rapidamente, ottenendo tuttavia variazioni di $V_{in,j}$ maggiori dell'errore di misura. La procedura di determinazione di A_j permette nominalmente di risolvere il problema, dato che il valore corrispondente risulta dal rapporto tra tensioni misurate “quasi” contemporaneamente.

Ho quindi eseguito un best-fit dei dati col metodo del minimo χ^2 usando la funzione Eq. 5 moltiplicata per una costante a , cioè la

$$A(f) = \frac{a}{\sqrt{1 + (f/f_T)^2}}, \quad (9)$$

e lasciando come parametri liberi di fit a e f_T .

Il best-fit riproduce in modo adeguato i dati sperimentali, come mostrato in Fig. 1, dove il pannello superiore riporta il grafico dei residui normalizzati (ho usato il complicato pacchetto `gridspace`, che si importa da `matplotlib`, per creare subplots di diversa altezza). I risultati del best-fit sono

$$a = 0.995 \pm 0.007 \quad (10)$$

$$f_T = (479 \pm 5) \text{ Hz} \quad (11)$$

$$\chi^2/\text{ndof} = 5.3/11 \quad (12)$$

$$\text{norm. cov.} = -0.89 \quad (13)$$

$$\text{absolute_sigma} = \text{False}. \quad (14)$$

Si nota che il parametro a è compatibile con l'unità, come atteso, e che f_T è compatibile con $f_{T,att}$ (grazie anche

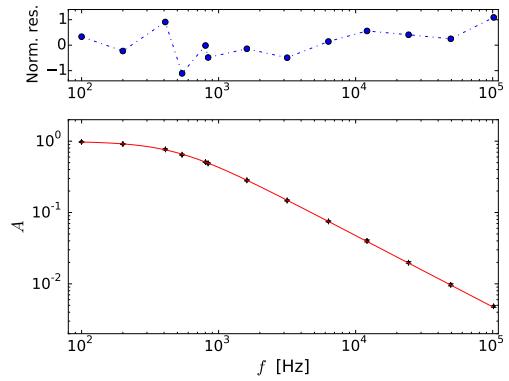


Figura 1. Dati sperimentali e risultato del best-fit a due parametri per le misure sul filtro passa-basso, eseguito come discusso nel testo. Il pannello superiore riporta il grafico dei residui normalizzati.

alla grande incertezza con cui il valore atteso è stato determinato). I due parametri di fit sono fortemente (anti)correlati l'un l'altro. Per questo motivo un best-fit imponendo $a = 1$ e lasciando come solo parametro libero f_T potrebbe risultare più sensato.

Si vede in ogni caso come esista, sia nei dati che nei risultati del modello, un'ampia regione di transizione, non limitata per alte frequenze, in cui il guadagno diminuisce in funzione della frequenza a significare il comportamento passa-basso del filtro costruito.

A. Misure sul filtro passa-alto

Ho fatto le stesse operazioni (misura, grafico, best-fit) anche per un filtro passa-alto, costruito con $R = (672 \pm 5)$ ohm e $C = 0.1 \mu\text{F}$ (con tolleranza del 10%). Per questo filtro mi aspetto $f_T = (2.4 \pm 0.3)$ kHz. La scelta della frequenza di taglio un po' più alta che non per il passa-basso aiuta a fare misure (qui sono interessanti i dati presi a $f < f_T$ e in questo modo si riesce a esplorare facilmente un buon intervallo di frequenze). Con la disponibilità di condensatori del laboratorio, questo comporta di usare una resistenza R relativamente prossima alla resistenza interna del generatore, per cui l'ipotesi che gli effetti relativi possano essere trascurati deve essere attentamente valutata.

I valori delle frequenze caratteristiche ottenuti dalle misure, risultati tutti in buon accordo con le attese, sono: $f_T = (2220 \pm 10)$ Hz [$\Delta\phi_T = (0.23 \pm 0.04) \pi$ rad], $f_{1/2} = (1280 \pm 12)$ Hz [$\Delta\phi_{1/2} = (0.35 \pm 0.03) \pi$ rad].

In analogia con quanto riportato nella sezione precedente, ho anche qui fatto determinare a Python l'attenuazione A_j e il corrispondente array di incertezze ΔA_j , ho disegnato il grafico in rappresentazione bilogaritmica e fatto un best-fit secondo la funzione Eq. 7 moltiplicata

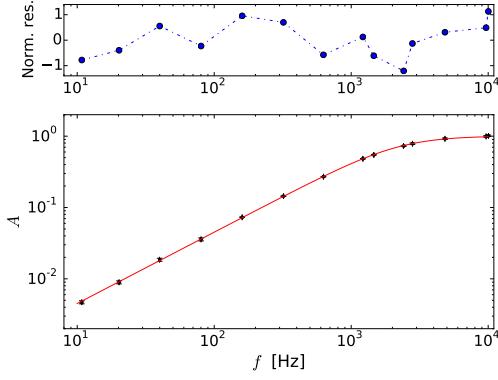


Figura 2. Dati sperimentali e risultato del best-fit a due parametri per le misure sul filtro passa-alto, eseguito come descritto nel testo. Il pannello superiore riporta il grafico dei residui normalizzati.

per il parametro a , cioè

$$A(f) = \frac{a}{\sqrt{1 + (f_T/f)^2}}, \quad (15)$$

lasciando come parametri liberi a e f_T .

Anche in questo caso il best-fit descrive in modo appropriato i dati sperimentali, come si vede in figura. I risultati sono

$$a = 1.003 \pm 0.008 \quad (16)$$

$$f_T = (2.23 \pm 0.03) \text{ kHz} \quad (17)$$

$$\chi^2/\text{ndof} = 6.3/12 \quad (18)$$

$$\text{norm. cov.} = 0.93 \quad (19)$$

$$\text{absolute_sigma} = \text{False}. \quad (20)$$

e per essi valgono le considerazioni generali svolte per il filtro passa-basso. Anche qui la forte correlazione tra i parametri potrebbe suggerire di eseguire un best-fit a un solo parametro, imponendo $a = 1$.

IV. ATTENUAZIONE DI -20 dB/DECADe

Nel caso di $f >> f_T$ il guadagno, o attenuazione, del filtro passa-basso si può approssimare attraverso sviluppo al primo ordine con $A(f) \sim f_T/f$. Analogamente per il passa-alto, nel caso di $f << f_T$, il guadagno, o attenuazione, si può approssimare con $A(f) \sim f/f_T$. Notate che questi andamenti sono “universali”, nel senso che, una volta stabilita la frequenza di taglio, essi valgono per qualsiasi filtro RC “a un polo”.

Prendiamo un passa-basso e misuriamone l'attenuazione a due diverse frequenze (tutte e due *superiori* a f_T), f_1 e $f_2 = 10f_1$: si avrà $A_2 = A_1/10$, cioè l'ampiezza del segnale in uscita alla frequenza f_2 è $(f_2/f_1) = 10$ volte più piccola rispetto a quella a frequenza f_1 . Che questo sia confermato dalle misure si vede benissimo in Fig. 1. Analogamente, nel caso del filtro passa-alto si vede come,

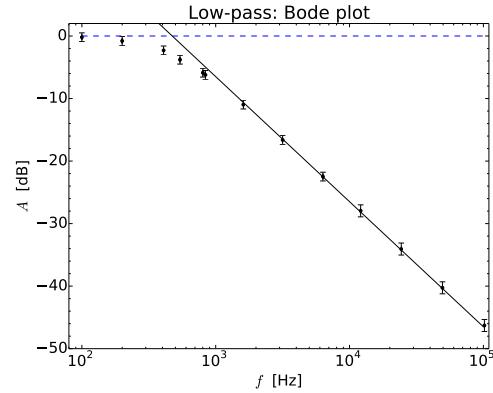


Figura 3. Diagramma di Bode e best-fit della regione di transizione per il filtro passa-basso, eseguito come descritto nel testo.

per frequenze *inferiori* a f_T , la frequenza più bassa (f_1) sia attenuata di un fattore (f_2/f_1) rispetto alla frequenza più alta (f_2).

Quando la frequenza viene moltiplicata per un fattore 10 si dice che essa è variata di una *decade*. Se si misura l'attenuazione in dB, a frequenze rispettivamente più alte o più basse di f_T (per passa-basso o passa-alto), si vede che, quando si varia la frequenza di una decade, la variazione dell'attenuazione in dB è $20 \log_{10}(1/10) = -20$ dB. In altre parole questi filtri *attenuano di -20 dB per decade*, ovvero la pendenza della *curva di risposta* è di -20 dB/decade. Poiché una decade è composta da un po' più di tre ottave, si dice anche che questi filtri *attenuano di -6 dB per ottava*.

A. Bode plot per il passa-basso

Passiamo ora al diagramma di Bode. Esso è una rappresentazione dell'attenuazione in dB di un filtro in funzione della frequenza (o, qualche volta, del rapporto f/f_T), cioè è un grafico, in scala *semilogaritmica* (l'asse logaritmico è l'*orizzontale*, quello delle frequenze), della grandezza $AdB_j = 20 \log_{10}(A_j)$. Per costruire il diagramma di Bode occorre quindi graficare un array che contiene tale grandezza [e un array che contiene l'incertezza da attribuire ai dati, costruito usando le regole di propagazione dell'errore come $\Delta AdB_j = (20/\ln(10))(\Delta A_j/A_j)$]. Il risultato è mostrato in Fig. 3.

Per aggiungere ancora un po' di pepe (o sale), ho deciso di eseguire un best-fit dei dati nella sola *regione di transizione*, quella in cui l'attenuazione segue l'andamento $A(f) \sim f_T/f$. Questa regione è quella che nel Bode plot ha un andamento lineare, così come era lineare l'andamento osservato in Fig. 1. Ho deciso arbitrariamente di considerare nel best-fit solo i dati acquisiti a $f > 1$ kHz, che corrisponde grosso modo al doppio della frequenza di taglio f_T determinata prima.

Per selezionare i punti da usare nel best-fit (solo nel best-fit, per la rappresentazione li voglio graficare tutti) ho costruito un array di supporto attraverso un ciclo `for` e una condizione sul valore f_j (la condizione è posta attraverso l'istruzione `if`, siete invitati a provare). Il best-fit di questi dati può essere eseguito sia in modo analitico (si tratta di un andamento lineare ed è consigliabile usare il calcolo analitico) che in modo numerico. Per semplicità, io ho usato il metodo numerico, che permette anche di ottenere l'espressione asintotica dell'errore sui parametri usando l'opzione `absolute_sigma = False` (le incertezze sulle misure non hanno carattere sicuramente o prevalentemente statistico).

I dati che intendo fissare sono espressi in dB, cioè essi sono già stati modificati per l'estrazione del logaritmo a base 10 (e poi moltiplicati per 20, secondo la definizione). Dunque la funzione di fit, che chiamerò qui $g(f)$, deve essere espressa in maniera congrua. Essa può infatti essere determinata calcolando il $20 \log_{10}(A(f))$, con $A(f) = f_T/f$, cioè

$$g(f) = 20 \log_{10} \left(\frac{f_T}{f} \right) = \quad (21)$$

$$= \kappa - 20 \log_{10} f, \quad (22)$$

con $\kappa = 20 \log_{10} f_T$ unico parametro libero di fit (si noti come nelle espressioni del logaritmo si intendano argomenti adimensionali, come se entrambe le frequenze fossero state divise per un valore comune di riferimento opportunamente dimensionato).

Ovviamente con questa scelta della funzione di best-fit occorre fornire un valore iniziale adeguato alla routine di minimizzazione e inoltre, a best-fit eseguito, occorre riconvertire il valore del parametro di best-fit e della sua incertezza ai valori del "parametro fisico" $f_T = 10^{\kappa/20}$ e della sua incertezza $\Delta f_T = (\Delta\kappa/20)10^{\kappa/20} \ln(10)$, dove $\Delta\kappa$ è l'incertezza sul parametro di fit.

La curva di best-fit è sovrapposta ai dati sperimentali (matematicamente trattati) in Fig. 3 e si nota un buon accordo. Il risultato del best-fit è

$$f_T = (473.4 \pm 6.2) \text{ Hz} \quad (23)$$

$$\chi^2/\text{ndof} = 1.1/6 \quad (24)$$

$$\text{absolute_sigma} = \text{False}. \quad (25)$$

che è ancora in accordo con le attese (ma notate come sia l'incertezza nella valutazione di f_T che il χ^2 siano aumentati, anche a causa dell'impiego di un set ridotto di dati e dell'uso di un singolo parametro di fit).

Come ultimo aspetto, il grafico di Fig. 3 mostra anche la cosiddetta *corner frequency* f_C , che corrisponde all'intersezione fra la curva che descrive l'attenuazione del filtro nella regione di transizione (la retta continua) e la retta $A [\text{dB}] = 0$ (linea tratteggiata in blu). Avendo eseguito il best-fit della regione di transizione, è facile individuare analiticamente tale intercetta: deve infatti essere $f_C = 10^{\kappa/20}$. Osservate che questa frequenza *corrisponde* di fatto alla frequenza di taglio f_T così come è

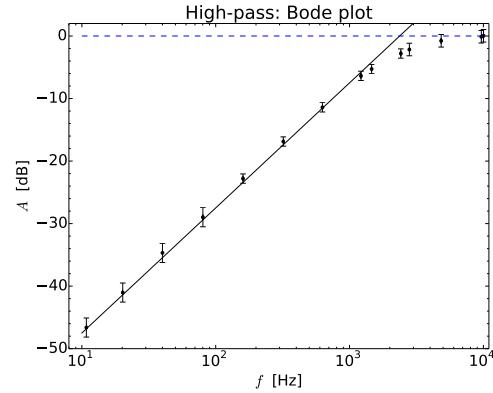


Figura 4. Diagramma di Bode e best-fit della regione di transizione per il filtro passa-alto, eseguito come descritto nel testo.

stata determinata sopra. Dunque $f_C = f_T$, come atteso per un filtro ben funzionante e ben descritto da un modello adeguato.

B. Bode plot per il passa-alto

Naturalmente il diagramma di Bode può essere costruito anche per il filtro passa-alto, usando lo stesso approccio messo a punto sopra e con la sola avvertenza di considerare, per il fit della regione di transizione, i dati corrispondenti a frequenze inferiori alla f_T (nell'esempio qui riportato, i dati sono quelli acquisiti per $f < 1.5 \text{ kHz}$).

Il Bode plot corrispondente è mostrato in Fig. 4. Il risultato del best-fit alla regione di transizione è

$$f_T = (2255 \pm 63) \text{ Hz} \quad (26)$$

$$\chi^2/\text{ndof} = 0.6/6 \quad (27)$$

$$\text{absolute_sigma} = \text{False}. \quad (28)$$

ancora in ragionevole accordo con le attese; infine la corner frequency f_C determinata graficamente risulta compatibile con f_T .

APPENDICE: NYQUIST PLOT

Nell'ambito dei grafici che servono a trattare il comportamento spettrale di circuiti e dispositivi, può essere rilevante citare brevemente un ulteriore metodo di rappresentazione che prende il nome, assieme a tanti altri metodi/fenomeni/modelli, da Harry Nyquist.

Per un sistema che ha funzione di trasferimento $T(f)$, si chiama *diagramma di Nyquist* il grafico parametrico $\text{Im}T(f)$ vs $\text{Re}T(f)$, ovvero il grafico costruito riportando sulle ordinate i valori della parte immaginaria e sulle ascisse quelli della parte reale della funzione di trasferimento calcolati per diverse frequenze. È evidente che un

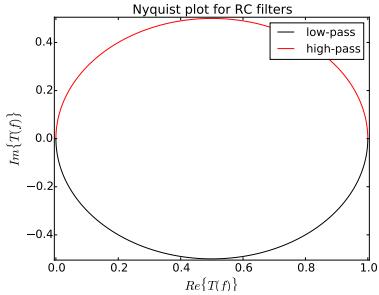


Figura 5. Diagramma di Nyquist costruito come descritto nel testo per i filtri RC passa-basso e passa-alto considerati nella nota. Nei grafici sono state considerate solo frequenze positive fino a $f = 0.1$ MHz.

grafico di questo tipo può essere costruito solo sulla ba-

- [1] Senza entrare nei dettagli, almeno per il momento, è intuitivo che ogni strumento di misura si comporta come filtro passa-banda nei confronti della grandezza misurata. Infatti nessuno strumento di misura reale è in grado di seguire variazioni istantanee (nessuno strumento reale ha una “prontezza” infinita!), per cui la misura è accurata solo per variazioni che avvengono in un certo intervallo temporale non nullo. Trasferendo l’approccio alla misura di segnali periodici, è ovvio che solo un certo intervallo di frequenze viene “correttamente” riprodotto nella misura. Uno dei parametri maggiormente rilevanti per caratterizzare la qualità di un oscilloscopio è la cosiddetta *banda passante*, che, come suggerito dal nome, rappresenta proprio l’ampiezza in frequenza della regione spettrale (cioè di frequenze) nella quale il segnale è ben riprodotto. Il pannello dello strumento impiegato in laboratorio riporta chiaramente tale banda passante (50 MHz): pensate a cosa può verificarsi quando il segnale in ingresso ha una frequenza maggiore rispetto a questo valore e anche a quanto vale l’estremo inferiore dell’intervallo di frequenze per le quali la misura è “corretta”, sia per accoppiamento in ingresso DC che AC.
- [2] Ho fatto poche misure: siete invitati a farne di più nelle vostre esperienze, privilegiando la *regione di transizione*. Infatti serve a poco aumentare il numero di misure alle frequenze che vengono lasciate passare dal filtro ed è “complicato” (e rumoroso) aumentare il numero di quelle alle frequenze che vengono fortemente attenuate.
- [3] Siete fortemente invitati a scrivere la funzione di trasferimento del filtro considerando anche r_G . Se lo farete correttamente, troverete che tale funzione è *identica* a quella di Eq. 3, per cui l’effetto della $r_G \neq 0$ è rigorosamente trascurabile. Infatti il segnale che indichiamo come V_{in} è misurato a valle della r_G , per cui questa non può avere effetto né nella determinazione di f_T , né nell’andamento di $T(f)$.
- [4] Il modello di oscilloscopio reale prevede la presenza che

se del modello (non esistono dati sperimentali che diano informazioni dirette su grandezze immaginarie!).

Per divertimento ho prodotto il diagramma di Nyquist relativo ai filtri passa-basso e passa-alto considerati in questa nota: il grafico è mostrato in Fig. 5. La rappresentazione della funzione di trasferimento tramite il Nyquist plot è rilevante per circuiti, o sistemi, con caratteristiche particolari, per esempio la presenza di divergenze o poli di cui eventualmente ci occuperemo brevemente in futuro. Nel caso dei filtri RC questa rappresentazione non è particolarmente significativa, poiché le caratteristiche fondamentali, in particolare la frequenza di taglio, non sono identificabili dalla lettura del grafico. Tuttavia può essere carino verificare qualche caratteristica del grafico, per esempio cercare dove è rappresentato il comportamento ad alta e bassa frequenza e dove quello alla frequenza di taglio. Provate a compiere questo tipo di analisi!

il segnale in ingresso all’oscilloscopio veda il parallelo di una resistenza $r_{osc} = 1$ Mohm e di un condensatore $C_{osc} = 25$ pF (valori nominali, riportati chiaramente anche sul pannello dello strumento). Dunque $Z_{osc} = r_{osc}/(1 + j\omega r_{osc}C_{osc})$. Nel range di frequenze esplorato si vede facilmente che $|Z_{osc}| \simeq r_{osc} > |Z_C|$, per cui, nell’accuratezza con cui ordinariamente si compiono misure con l’oscilloscopio, la sua presenza in parallelo al condensatore C produce effetti trascurabili. Siete fortemente invitati a compiere le opportune verifiche numeriche e a tenere presente che le affermazioni riportate si riferiscono, ovviamente, all’accoppiamento in DC del canale dell’oscilloscopio.

- [5] Può essere oggetto di significative discussioni cercare validi motivi per evitare di largheggiare eccessivamente nella determinazione dell’incertezza su A_j . Può essere ragionevole supporre che sia le incertezze di calibrazione che quelle di lettura siano indipendenti per i due canali, per cui la somma potrebbe essere eseguita in quadratura. Supponendo trascurabili i contributi delle incertezze di lettura (ipotesi abbastanza ragionevole nel caso in cui le tracce siano molto sottili e, soprattutto, coprano gran parte dello schermo dell’oscilloscopio), l’errore relativo su A_j passerebbe dal 6% a circa il 4%. Inoltre, nell’ulteriore ipotesi che l’errore di calibrazione sia costante nel tempo (ipotesi non molto ragionevole in presenza di strumenti che producono molto calore, come l’oscilloscopio) e che l’ampiezza V_{in} non cambi tra una misura e l’altra in modo tale da richiedere un cambio di scala, l’incertezza di calibrazione su V_{in} potrebbe essere considerata come un *errore di scala* nell’asse verticale dei grafici, e i punti lì rappresentati potrebbero limitarsi a mostrare barre di errore dovute solo all’incertezza sulla misura di V_{out} (qui è certamente necessario impiegare diverse scale se si vuole rendere ragionevolmente piccolo l’errore di lettura). Queste “manipolazioni” dell’errore dipendono fortemente dalla validità delle varie assunzioni e devono sempre essere accompagnate da motivate giustificazioni!

Segnali periodici non alternati integrati nel tempo

francesco.fuso@unipi.it

(Dated: version 1 - Francesco Fuso, 21 novembre 2021)

Questa nota riporta qualche breve discussione sull'uso di un integratore con forme d'onda non alternate, in particolare onde quadre positive e con un duty cycle variabile.

I. INTRODUZIONE

L'esperimento descritto in questa nota, che fa parte di un'esercitazione articolata su diversi punti, prevede di inviare un'onda quadra positiva (livello basso circa 0 V, livello alto circa 5 V) e asimmetrica (duty cycle variabile) a un integratore RC.

A. Integratore RC

In questa sezione richiamiamo brevemente il comportamento di un integratore RC, ovvero della serie di una resistenza R e un condensatore C in cui il segnale in uscita $V_{out}(t)$ viene preso ai capi del condensatore (che si trova quindi "in parallelo" al segnale), come mostrato in Fig. 1. Supponendo trascurabile la corrente che fluisce fuori dalla serie per andare, ad esempio, allo strumento di misura, se l'ingresso $V_{in}(t)$ è una forma d'onda quadra (periodica) il condensatore compie periodicamente cicli di carica e scarica; in particolare, poiché la forma d'onda è prodotta da un generatore *reale* di d.d.p., modellato attraverso la serie di un generatore ideale e di una resistenza interna (o di Thévenin) r_G , i processi di carica e scarica avvengono attraverso la maglia che comprende la serie $r_G + R$, per cui la costante tempo caratteristica è $\tau = (r_G + R)C$. Carica e scarica sono modellate con le note funzioni esponenziali del tempo:

$$V_{out}(t) \propto 1 - \exp[-(t - t_{0C})/\tau] \quad \text{carica} \quad (1)$$

$$V_{out}(t) \propto \exp[-(t - t_{0S})]/\tau \quad \text{scarica ,} \quad (2)$$

dove t_{0S} , t_{0C} rappresentano gli istanti in cui hanno inizio le fasi rispettivamente di carica e scarica. Notate che le funzioni appena scritte devono essere raccordate fra loro, per cui le costanti di proporzionalità per ogni fase dipendono dalla fase precedente.

Come abbiamo facilmente verificato in laboratorio, e come è ovvio, il raggiungimento delle condizioni asintotiche dipende dal tempo lasciato a disposizione del condensatore per caricarsi e scaricarsi. Questo tempo vale $T/2$, cioè metà del periodo della forma d'onda, e gli asintoti di carica e scarica possono essere raggiunti solo se $T > \tau$.

Nelle condizioni opposte, cioè per $T \ll \tau$, solo un tratto delle fasi di carica e scarica può essere compiuto; in termini matematici, le funzioni esponenziali tendono a linearizzarsi in accordo con uno sviluppo di Taylor al primo ordine. Per esempio, nel caso della carica, supponendo $t_{0C} = -T/4$ (la fase di carica termina all'istante $T/4$,

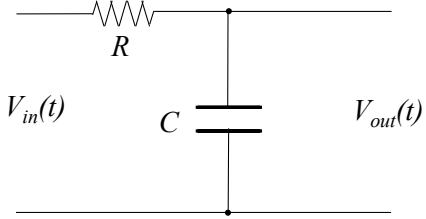


Figura 1. Schema del semplice circuito integratore considerato nel testo.

per cui la sua durata complessiva è $T/2$) e sviluppando attorno a $t = 0$, si ha

$$1 - \exp[-(t - t_{0C})/\tau] \simeq t/\tau , \quad (3)$$

dove i termini trascurati nella serie sono di ordine $(t/\tau)^2$, o superiore; tenendo conto che, al massimo, $t = T/4$, l'errore relativo che si compie nella linearizzazione ha ordine T/τ (con dei coefficienti davanti che sono minori di uno), tanto più trascurabile quanto più T è piccolo rispetto a τ . L'Eq. 3 rappresenta un andamento lineare con t per la fase di carica e, come si può facilmente dimostrare, la fase di scarica, linearizzata, conduce a un andamento proporzionale a $-t/\tau$. Poiché le varie fasi devono essere raccordate tra loro, $V_{out}(t)$ tende ad assumere una forma d'onda triangolare. In termini qualitativi, questa forma d'onda può essere considerata come l'*integrale nel tempo* della $V_{in}(t)$ in ingresso (qudra). Ovviamente, il cambio di forma d'onda tra ingresso e uscita è accompagnato da una cospicua riduzione dell'ampiezza, o ampiezza picco-picco, che è anche dell'ordine di T/τ (a parte coefficienti).

La riduzione dell'ampiezza può anche essere interpretata in modo diverso: il condensatore non fa mai in tempo ad accumulare una sufficiente quantità di carica sulle sue armature, per cui l'intensità di corrente I che fluisce nel circuito per giungere al condensatore si mantiene mediamente alta. La caduta di potenziale ai capi della resistenza, proporzionale all'intensità di corrente che la attraversa, è anche rilevante, per cui l'ampiezza del segnale in uscita, V_{out} , si riduce rispetto a quella, V_{in} , in ingresso. Questo ragionamento può essere formalizzato come segue:

$$RI = V_{in} - V_{out} \simeq V_{in} , \quad (4)$$

dove abbiamo appunto supposto $V_{out} \ll V_{in}$. D'altra parte la corrente è fatta di cariche che vanno sull'armatura del condensatore, per cui la sua intensità può essere

espressa come $I(t) = dQ(t)/dt$, dove, ancora una volta, trascuriamo la corrente che fluisce fuori dal circuito. Infine, notando che $V_{out}(t) = Q(t)/C$, si ottiene, sostituendo al primo ordine $\simeq \text{con} = \text{in}$ Eq. 4,

$$R \frac{dQ(t)}{dt} = RC \frac{dV_{out}(t)}{dt} = V_{in}(t); . \quad (5)$$

L'ultima uguaglianza vale istante per istante e dunque la stessa relazione vale anche integrando nel tempo entrambi i membri:

$$RC \int \frac{dV_{out}(t)}{dt} dt = \int V_{in}(t') dt', \quad (6)$$

che formalmente equivale a

$$V_{out}(t) = \frac{1}{RC} \int V_{in}(t') dt'. \quad (7)$$

Questa equazione dimostra che il segnale in uscita può approssimare (al primo ordine) l'integrale di quello di ingresso, da cui la denominazione di *integratore* che si dà al circuito. L'approssimazione è tanto più accurata quanto più il segnale in uscita risulta attenuato rispetto all'ingresso: a prescindere dalla forma d'onda impiegata in ingresso e, più in generale, dall'andamento temporale del segnale di ingresso (che potrebbe ovviamente anche essere non periodico), il segnale di uscita è tanto più attenuato quanto più il condensatore si mantiene scarico, cioè quanto più il tempo caratteristico τ del circuito risulta lungo rispetto al tempo scala di variazione del segnale in ingresso.

L'integrale di Eq. 7 è chiaramente calcolato in modo “analogo”, per cui non è possibile stabilire con precisione quali siano gli estremi di integrazione. L'estremo superiore è in realtà coincidente con l'istante t nel quale si misura V_{out} ; per l'estremo inferiore si può ritenere che esso preceda t di un intervallo dell'ordine di τ . Più precisamente, tenendo conto che in un andamento esponenziale il tempo caratteristico rappresenta una frazione del tempo necessario per raggiungere le condizioni asintotiche, l'estremo inferiore di integrazione potrebbe essere espresso come t meno qualche τ , dove *qualche* significa, convenzionalmente, 3 – 5. Potete facilmente rendervene conto supponendo che $V_{in}(t)$ sia una funzione a gradino: una volta completata l'integrazione, l'integrale in Eq. 7 dovrebbe risultare praticamente pari all'ampiezza del gradino (moltiplicata per il tempo che si è atteso) e perché questo si verifichi occorre attendere un tempo pari a qualche τ .

B. Integrale e media temporale

Anche se la denominazione integratore è comunemente accettata per il circuito di Fig. 1, ovviamente impiegato nelle condizioni espresse in precedenza ($T \ll \tau$ nel caso di forme d'onda periodiche), una semplice considerazione dimensionale sull'Eq. 7 mostra che la funzione del

circuito è piuttosto quella di fornire l'integrale nel tempo di $V_{in}(t)$ debitamente diviso per la costante tempo RC (essa equivale a τ nel caso, praticamente frequente, $r_G \ll R$). L'operazione risultante somiglia quindi a una *media temporale*, eseguita ovviamente in modo analogico, cioè su un intervallo temporale che non può essere definito con precisione, ma che vale *qualche* τ .

Se la $V_{in}(t)$ in ingresso è una forma d'onda *alternata*, cioè a media nulla sul suo periodo T , nel caso $T \ll \tau$ l'ampiezza V_{out} tende a zero; infatti l'integrale compiuto su un numero intero di cicli fa zero, e la parte rimanente, come si può facilmente verificare, è dell'ordine di $T/(2RC)$. Se $T \ll \tau = (r_G + R)C$, nella situazione praticamente frequente in cui $r_G \ll R$ è anche $T \ll RC$, per cui, a prescindere dal fatto che l'integrale non viene calcolato su un numero intero di periodi, si ha sempre $V_{out} \rightarrow 0$.

Questa circostanza illustra uno dei mille impieghi pratici che può avere il circuito di Fig. 1: supponete infatti di dover misurare un segnale che varia su un tempo scala “sufficientemente lungo” e a cui è sovrapposto del rumore, costituito da fluttuazioni statistiche a media temporale nulla, e magari anche dei segnali spuri *alternati* e periodici, di frequenza “sufficientemente alta”. Usando un integratore opportunamente dimensionato sarete in grado di *ripulire* il segnale, abbattendo il rumore statistico e anche il contributo di segnale spurio. Questo presupone ovviamente che la misura “integrata” venga compiuta in un tempo sufficientemente lungo (*qualche* τ) e anche, per ovvi motivi, che il segnale di vostro interesse possa essere considerato costante in questo intervallo di tempo.

Integrare temporalmente il segnale è in effetti una delle strategie più comuni per aumentare il rapporto segnale/rumore nel caso di segnali costanti o lentamente variabili nel tempo. Accenneremo in futuro a tecniche un po' più raffinate che permettono di usare un approccio concettualmente simile anche per segnali che variano non troppo lentamente nel tempo.

Accanto a questo impiego nell'ambito della misura, un integratore RC può sicuramente trovarne altri. È infatti evidente che un segnale in ingresso $V_{in}(t)$ periodico ma *non alternato* tende al proprio valore medio quando processato dall'integratore. Questa è la strategia generalmente impiegata quando si vuole rendere continuo (cioè costante), o meglio *quasi-continuo*, un segnale oscillante, anche alternato: tale segnale viene prima reso non alternato e quindi inviato all'ingresso di un qualche integratore, spesso di tipo RC. Come tutti sapete (spero!), la rete elettrica fornisce una d.d.p. periodica (a frequenza 50 Hz), di forma nominalmente sinusoidale e alternata e di ampiezza relativamente alta (230 – 240 V_{rms}). Nei comuni alimentatori per dispositivi elettronici (computer, telefoni e tutto quello che può venirvi in mente), dove è necessaria una d.d.p. continua e di ampiezza relativamente bassa (qualche V), esistono dispositivi basati su *diodi* che permettono di rendere non alternata la d.d.p. (opportunamente diminuita in ampiezza da altri dispositivi, basati generalmente su *trasformatori*). Essa viene ul-

riormente *livellata* usando proprio degli integratori. Di tutto questo avremo modo di occuparci nel prosieguo del corso: per ora limitatevi a osservare che lo stesso approccio può essere utilizzato negli strumenti di misura che forniscono una lettura rms dell'ampiezza di segnali alternati, per esempio i multimetri utilizzati come voltmetri in alternata.

II. INTEGRAZIONE DI SEGNALI CON DUTY CYCLE VARIABILE

Esiste sicuramente un ulteriore impiego dell'integratore rilevante per i nostri scopi sperimentali. Arduino, come sapete, è un onesto digitalizzatore, cioè un dispositivo che converte una d.d.p. analogica (virtualmente variabile in modo continuo) in una misura digitale, cioè un numero intero (di unità arbitrarie di digitalizzazione, ovvero digit nel nostro linguaggio). Quindi esso si comporta da onesto convertitore A/D (analogico/digitale). L'operazione opposta, quella di convertire un numero in una d.d.p., è più complicata e non è implementata nella versione di Arduino (Uno) da noi usata attualmente.

Esiste però un escamotage, che troveremo molto utile per realizzare alcuni esperimenti. Infatti Arduino è in grado di produrre su alcune delle sue uscite digitali (quelle marcate con un tilde negli schemi e nelle serigrafie applicate alla scheda) un'onda quadra con *duty cycle variabile* secondo un numero intero (che può andare da 0 a 255, ma di questi dettagli tratteremo in seguito). Questa onda quadra asimmetrica, che ha una frequenza attorno a 1 kHz ed è già di per sé non alternata (Arduino non può fornire d.d.p. inferiori al valore della linea di riferimento, cioè di massa o terra), può essere inviata a un integratore opportunamente dimensionato, da cui essa esce come un segnale quasi-continuo la cui ampiezza dipende linearmente dal numero impiegato per determinare il duty cycle. Tutto l'insieme fornisce quindi un modo per creare una d.d.p. quasi-continua di ampiezza (positiva) determinata digitalmente via software.

L'esperimento qui descritto è un'illustrazione di quanto appena descritto. In esso l'onda quadra asimmetrica è prodotta dal generatore di forme d'onda e Arduino è impiegato come digitalizzatore, così come siamo abituati a fare. Lo schema del circuito è mostrato in Fig. 2: si vede come Arduino registri entrambi i segnali $V_{in}(t)$ e $V_{out}(t)$ in ingresso e uscita dall'integratore RC, segnali che vengono anche monitorati sui due canali dell'oscilloscopio. Infatti Arduino ha la possibilità di campionare e digitalizzare più segnali (fino a un massimo di 8). Purtroppo l'impiego contemporaneo di più porte analogiche di ingresso (A0 e A2, nel nostro caso) impedisce di utilizzare gli accorgimenti di "overclock" generalmente impiegati in laboratorio e, più in generale, peggiora l'accuratezza di timing del campionamento. Di conseguenza è possibile che l'intervallo temporale Δt tra un campionamento e il successivo sia non costante entro l'incertezza convenzionale (4 μs), ma questo è irrilevante per gli scopi dell'esperienza.

Inoltre la necessità di registrare all'interno della memoria di Arduino due array, corrispondenti alle letture digitalizzate delle due porte, impone lunghezze dei record limitate: nello specifico, la combinazione di sketch e script (nome comune *duty2021*) riportata in Appendice conduce a record costituiti da 128 righe e tre colonne (rispettivamente tempo in μs , V_{in} e V_{out} in digit). La durata complessiva dell'acquisizione è di circa 100 ms, tempo piuttosto lungo che impone di operare a frequenze f della forma d'onda quadra in ingresso mantenute sufficientemente basse (qualche decina di Hz).

Nell'esperimento realizzato in pratica si è scelta $f = (57.26 \pm 0.02)$ Hz, corrispondente a $T = (17.46 \pm 0.06)$ ms, e si è posto $\tau = (32 \pm 3)$ ms, con incertezza dominata dalla tolleranza con cui è nota la capacità C . Le condizioni che si realizzano sono dunque quelle in cui $T < \tau$, diverse da quelle ordinariamente scelte, e discusse in precedenza, in cui $T \ll \tau$. Ci si aspetta quindi che il processo di media temporale non sia completo e che $V_{out}(t)$ mantenga una variazione temporale di forma quasi-triangolare.

I segnali acquisiti per tre distinte regolazioni della manopola che regola il duty cycle nel generatore di funzioni (rispettivamente al minimo possibile, attorno al punto di mezzo e al massimo possibile) sono mostrati in Fig. 3; essi mostrano come il valore medio di $V_{out}(t)$, \bar{V}_{out} , si modifichi al variare del duty cycle. In particolare, quando il duty cycle è tale che la durata dell'onda quadra a livello alto (il tempo disponibile per la fase di carica del condensatore) è al minimo (secondo il manuale del generatore questa regolazione corrisponde a un duty cycle 20%:80%, con ovvio significato), \bar{V}_{out} è al suo minimo, e viceversa. In ogni caso si osserva come sia sempre $\bar{V}_{out} < \bar{V}_{in}$; infatti il tempo τ su cui viene effettuata l'integrazione, ovvero calcolata la media temporale, non è sufficiente per livellare completamente il segnale in ingresso. Infatti $V_{out}(t)$ mantiene un'oscillazione attorno al proprio valore medio, la cui forma è approssimativamente triangolare in accordo con le aspettative. Osservate come la deviazione dalla forma triangolare sia più marcata nelle fasi in cui il tempo a disposizione è maggiore: per intenderci, nel grafico del pannello in alto, dove il duty cycle è tale che il tempo disponibile per la carica è maggiore di quello per la scarica, la fase di carica mostra apprezzabili deviazioni dall'andamento lineare. Essi sono dovuti al contributo dei termini di ordine superiore nell'Eq. 3, che rendono l'andamento simile a quello di Eq. 4 ("esponenziale").

A. Ripple e tempi caratteristici

È evidente che, nelle condizioni $T \ll \tau$, ovvero, idealmente, $\tau \rightarrow \infty$, $V_{out}(t)$ raggiungerebbe a regime a un valore costante dipendente dal duty cycle della forma d'onda in ingresso. Come già accennato, livellare in questo modo segnali periodici non alternati è alla base della realizzazione degli alimentatori in "corrente continua". In quel contesto la discrepanza tra le condizioni ideali e quelle attuali di funzionamento è valutata attraverso

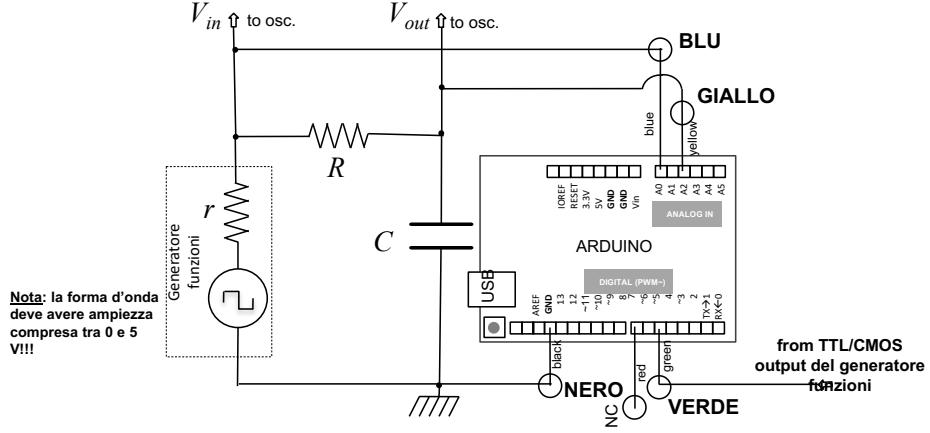


Figura 2. Schema del circuito per la digitalizzazione con Arduino di onde quadre asimmetriche e positive integrate temporalmente dal circuito RC.

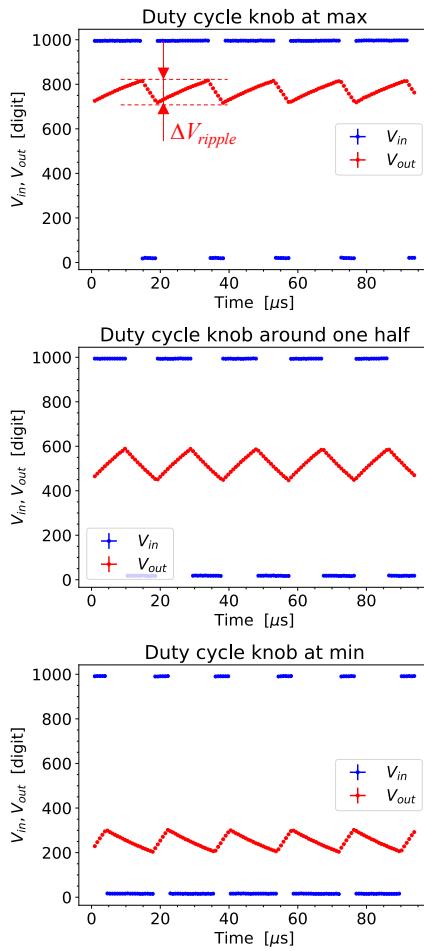


Figura 3. Segnali $V_{in}(t)$ e $V_{out}(t)$ acquisiti nel circuito di Fig. 2 nelle condizioni descritte nel testo. I grafici riportano l'incertezza convenzionale attribuita alle misure di Arduino, rispettivamente ± 1 digit e $\pm 4 \mu s$ per le ampiezze e i tempi.

una grandezza, $\Delta V_{ripple} = V_{max} - V_{min}$, che rappresenta la d.d.p. tra valore massimo (V_{max}) e valore minimo (V_{min}) assunto periodicamente da $V_{out}(t)$: per referenza, ΔV_{ripple} è indicato nel pannello in alto di Fig. 3. Una figura di merito rilevante per gli alimentatori è proprio l'ampiezza di ΔV_{ripple} , o anche il suo rapporto con il valore medio \bar{V}_{out} .

Supponendo che le condizioni che conducono alla linearizzazione espressa in Eq. 3 siano pienamente soddisfatte, $V_{out}(t)$ può essere espressa come

$$V_{out}(t) \simeq V_{min} \left(1 + \frac{t - t_{0C}}{\tau} \right) \quad \text{carica} \quad (8)$$

$$V_{out}(t) \simeq V_{max} \left(1 - \frac{t - t_{0S}}{\tau} \right) \quad \text{scarica} , \quad (9)$$

che al ripetersi periodico degli istanti di inizio carica e scarica, rispettivamente t_{0C} e t_{0S} , rappresenta una forma d'onda triangolare oscillante attorno a un valore medio diverso da zero. Detti Δt_C e Δt_S gli intervalli temporali delle fasi di carica e scarica, cioè le durate dei livelli alto e basso della forma d'onda quadra asimmetrica inviata in ingresso, si ha

$$\tau \simeq \Delta t_C \frac{V_{min}}{\Delta V_{ripple}} \quad \text{carica} \quad (10)$$

$$\tau \simeq \Delta t_S \frac{V_{max}}{\Delta V_{ripple}} \quad \text{scarica} . \quad (11)$$

Operando su una delle fasi di carica e di scarica dei dati ripotati nel pannello superiore di Fig. 3 si ottiene $\tau = (36 \pm 6)$ ms per la fase di scarica e $\tau = (107 \pm 5)$ ms per quella di carica, con incertezze dominate dall'intervallo di campionamento di Arduino ($\Delta t \simeq 0.7$ ms) usato come errore nella valutazione di Δt_C e Δt_S : si vede come il tempo caratteristico valutato per la fase di scarica sia in accordo con le aspettative, mentre lo stesso non si verifica per la fase di carica a causa della deviazione dal comportamento linearizzato; un best-fit secondo l'andamento di

Eq. 4 consente infatti di ritrovare il tempo caratteristico τ atteso anche in questo caso.

APPENDICE: SKETCH DI ARDUINO E SCRIPT DI PYTHON

Per l'esercitazione pratica si usa una combinazione di sketch e script, di nome duty2021 (con estensione .ino e .py per sketch e script, rispettivamente), che sono disponibili nel sito di e-learning in rete e riportati qui nel seguito.

```
/*
Questo sketch serve per acquisire 2 segnali (porte A0 e A2 di Arduino) con intervallo temporale prefissato
50 us nominali. L'acquisizione e' sincrona con il segnale TTL proveniente dal generatore di forme d'onda,
che viene inviato alla porta digitale pin5 di Arduino.
Lo sketch, usato in combinazione con il proprio script,
produce files di tre colonne, rispettivamente tempo in us,
d.d.p. su porta A0 in digit, d.d.p. su porta A2 in digit.
*/

// Blocco definizioni
const unsigned int analogPin0=0; // Definisce la porta A0 per la lettura
const unsigned int analogPin2=2; // Definisce la porta A2 per la lettura
const unsigned int sincPin = 5; //pin 5 ingresso digitale per la sincronizzazione con il generatore
int i; // Definisce la variabile intera i (contatore)
int delays = 500; // Definisce la variabile intera delays e la pone pari a 50 us
unsigned long StartTime; // Definisce la variabile StartTime
int V0[128]; // Definisce l'array intero V0 che contiene le misure sulla porta A0
int V2[128]; // Definisce l'array intero V2 che contiene le misure sulla porta A2
long t[128]; // Definisce l'array t che contiene i tempi di campionamento
int start=0; // Definisce il valore start (usato come flag)
int sinc; // Variabile di sincronizzazione
int dummy; // Variabile dummy letta inizialmente

// Istruzioni di inizializzazione
void setup()
{
    Serial.begin(19200); // Inizializza la porta seriale a 19200 baud
    Serial.flush(); // Pulisce il buffer della porta seriale
    pinMode(sincPin, INPUT); //pin sincPin configurato come ingresso digitale
    //bitClear(ADCSRA,ADPS0); // Istruzioni necessarie per velocizzare
    //bitClear(ADCSRA,ADPS2); // il rate di acquisizione analogica
    digitalWrite(7,HIGH); // Pone pin7 a livello alto (serve per calibrazione alternativa)
}

// Istruzioni del programma
void loop()
{
    if (Serial.available() >0) // Controlla se il buffer seriale ha qualcosa
    {
        dummy = (Serial.read()); // Legge lo start
        Serial.flush(); // Svuota la seriale
    start=1; // Pone il flag start a uno
    }

    if(!start) return // Se il flag e' start=0 non esegue le operazioni qui di seguito
                    // altrimenti le fa partire (quindi aspetta di ricevere l'istruzione
                    // di partenza prima di partire)
    delay(1000); // Aspetta 1000 ms per evitare casini
}
```

```

sinc = digitalRead(sincPin);//legge sincPin
while (sinc==LOW) // ciclo di attesa iniziale per sincronizzazione, attende che sincPin vada basso
{sinc = digitalRead(sincPin);} //legge sincPin
while (sinc==HIGH) // ciclo di attesa iniziale per sincronizzazione, attende che sincPin vada alto
{sinc = digitalRead(sincPin);} //legge sincPin
StartTime=micros(); // Misura il tempo iniziale con l'orologio interno
delayMicroseconds(delays);
for(i=0;i<2;i++) // Esegue due misure iniziali e poi non le registra (ci riscrive sopra) per minimizzare i
{
    V0[i]=analogRead(analogPin0);
    V2[i]=analogRead(analogPin2);
}
for(i=0;i<128;i++) // Loop di misura
{
    t[i]=micros()-StartTime; // Legge il timestamp e lo mette in array t
    V0[i]=analogRead(analogPin0); // Legge analogPin0 e lo mette in array V0
    V2[i]=analogRead(analogPin2); // Legge analogPin2 e lo mette in array V2
    delayMicroseconds(delays); // Aspetta tot us
}
for(i=0;i<128;i++) // Loop per la scrittura su porta seriale
{
    Serial.print(t[i]); // Scrive t[i]
    Serial.print(" "); // Mette uno spazio
    Serial.print(V0[i]); // Scrive V0[i]
    Serial.print(" "); // Mette uno spazio
    Serial.println(V2[i]); // Scrive V2[i] e va a capo
}
delay(1000); // Aspetta 1000 ms per evitare casini

start=0; // Annulla il flag
Serial.flush(); // Pulisce il buffer della porta seriale (si sa mai)
}

```

```

# Questo script serve per interfacciarsi con Arduino nell'esperienza
# dell'acquisizione sincrona di due canali (onda quadra con duty cycle)
# L'interfacciamento avviene attraverso:
# 1. scrittura di un carattere (byte)
# 2. lettura dei dati disponibili su porta seriale

import serial # libreria per gestione porta seriale (USB)
import time # libreria per temporizzazione
import pylab
import numpy

Directory='../../dati_arduino/' # nome directory dati
                                # << DA CAMBIARE SECONDO NECESSITA'
FileName=Directory+'dataduty.txt' # nome file << DA CAMBIARE SECONDO NECESSITA'

ard=serial.Serial('/dev/ttyACM0',19200) # apre porta seriale (occhio alla sintassi, dipende
                                         # dal sistema operativo!)
time.sleep(2) # aspetta due secondi per evitare casini

```

```
ard.write(b'1') # scrive un carattere sulla porta seriale che serve per far partire # l'acquisizione
time.sleep(1) # aspetta un secondo per evitare casini
outputFile = open(fileName, "w" ) # apre file dati carica per scrittura
print ("start")

# loop lettura dati da seriale (128 punti)
for i in range (0,128):
    data = ard.readline().decode() # legge il dato e lo decodifica
    if data:
        outputFile.write(data) # scrive i blocchi di dati

outputFile.close() # chiude il file dei dati di carica
ard.close() # chiude la comunicazione seriale con Arduino
print('end') # scrive sulla console che ha finito

t,V0,V2=pylab.loadtxt(fileName,unpack=True)

pylab.errorbar(t,V0,1,4,linestyle='-.',color='blue',marker='.')
pylab.errorbar(t,V2,1,4,linestyle='-.',color='red',marker='.')
pylab.rc('font',size=14)
pylab.xlabel('Time [$\mu$s]',size=16)
pylab.ylabel('$V_0, V_2$ [digit]',size=16)

pylab.show()
```

Curva caratteristica del diodo con Arduino

francesco.fuso@unipi.it; <http://www.df.unipi.it/~fuso/dida>

(Dated: version 10 - Lara Palla e Francesco Fuso, 9 dicembre 2016)

Questa nota discute alcuni aspetti di interesse per l'esperienza di registrazione della curva caratteristica I-V di un diodo bipolare a giunzione p-n condotta in laboratorio usando Arduino.

I. INTRODUZIONE

L'esperienza considerata in questa nota è piuttosto semplice dal punto di vista concettuale. Arduino è usato come scheda di input/output (I/O) in una modalità che permette di automatizzare la presa dati e di raccogliere un numero sufficiente di punti per ricostruire con buon dettaglio una curva sperimentale. La curva in questione è la cosiddetta *curva caratteristica I-V* di un diodo a giunzione bipolare (p-n) in silicio.

Un diodo a giunzione si comporta come un componente che ha una risposta decisamente non ohmica. Infatti la corrente (di intensità I) che attraversa la giunzione non è linearmente proporzionale alla differenza di potenziale (ΔV) applicata ai terminali del componente. Secondo il modello di Shockley, la legge che descrive il comportamento è

$$I = I_0 \left[\exp \left(\frac{\Delta V}{\eta V_T} \right) - 1 \right], \quad (1)$$

con I_0 corrente di saturazione inversa (del valore tipico dell'ordine di $1 - 10$ nA per i diodi usati in laboratorio, dunque molto piccola), η parametro costruttivo di valore tipico $\eta \simeq 1.5 - 2$ per gli ordinari diodi al silicio, V_T differenza di potenziale, talvolta definita *termica*, legata alla temperatura di operazione T , alla carica elementare e e alla costante di Boltzmann k_B attraverso la relazione $eV_T = k_B T$; poiché $k_B T \simeq 1/40$ eV a temperatura ambiente, si ha $V_T \simeq 26$ mV.

Tracciare la curva in questione richiede, in una semplice configurazione sperimentale, di poter disporre di un generatore di d.d.p. variabile, di misurare il valore effettivo ΔV applicato e di misurare la corrispondente intensità di corrente I che fluisce nel diodo. Tale semplice configurazione sperimentale è descritta schematicamente in Fig. 1(a) in cui si suppone implicitamente di poter trascurare gli effetti di tutte le resistenze interne (quella del generatore e del misuratore di corrente, ritenute trascurabili, e quella del misuratore di d.d.p., ritenuta così grande da non sottrarre corrente al resto del circuito).

In laboratorio non disponiamo di un generatore di d.d.p. variabile: potremmo realizzarne facilmente uno, per esempio costruendo un partitore di tensione con una resistenza variabile (potenziometro). Tuttavia, per avere una ricostruzione significativa della curva I-V [rappresentata per esempio in Fig. 1(b)] occorre registrare i dati su un numero relativamente elevato di punti corrispondenti a piccole variazioni del valore di ΔV , cosa non semplice dal punto di vista pratico.

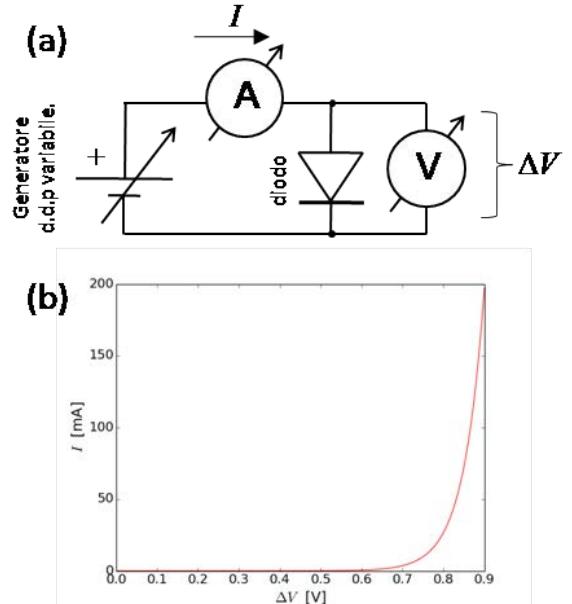


Figura 1. Schema concettuale di un'esperienza di ricostruzione della curva caratteristica I-V di un diodo (a) e curva calcolata secondo Eq. 1, supponendo $\eta = 2$, $V_T = 26$ mV e $I_0 = 3$ nA: per chiarezza è mostrato il solo ramo corrispondente a valori positivi della d.d.p. applicata, nel solo intervallo $\Delta V = 0 - 0.9$ V.

La scheda Arduino, opportunamente programmata e con l'aggiunta di pochi elementi circuituali esterni, può permettere un'acquisizione *automatizzata* via computer, cioè ottenere un file contenente un numero di punti sperimentali sufficiente per le ulteriori analisi (grafico, best-fit). Gli ingredienti necessari sono:

1. ottenere una d.d.p. variabile, ovvero una rampa di tensione che evolve nel tempo con tanti piccoli gradini;
2. registrare il valore della d.d.p. applicata al diodo per ogni gradino della rampa in modo automatico;
3. registrare il valore della corrispondente intensità di corrente che circola nel diodo; poiché le porte analogiche di Arduino consentono solo misure di d.d.p., questo punto richiede di "convertire" l'intensità di corrente in una opportuna tensione, cosa che può facilmente essere realizzata a posteriori, cioè lavorando sui dati grezzi acquisiti, sfruttando la legge di Ohm.

II. LA MODALITÀ PWM DI ARDUINO

Abbiamo più volte fatto riferimento ad Arduino come a una scheda I/O. Nell'esperienza che vogliamo svolgere ci sono sicuramente delle grandezze analogiche *in ingresso* (input) a Arduino che vogliamo leggere, cioè digitalizzare e acquisire con il computer. Queste grandezze analogiche sono quelle rappresentative di ΔV e I . Ci piacerebbe molto avere anche una grandezza analogica (variabile in maniera quasi continua e controllata) *in uscita* (output), cioè la d.d.p. che deve essere applicata al diodo.

Purtroppo, dal punto di vista tecnologico ottenere una grandezza analogica in uscita da una scheda I/O è più complicato che non eseguire la digitalizzazione di una grandezza analogica in ingresso. Il microcontroller di Arduino, infatti, non ha la possibilità di creare una d.d.p. di valore determinato a partire da un'istruzione digitale.

Arduino ha invece la possibilità di accendere o spegnere delle porte digitali in uscita, cioè di porle a potenziale nullo (entro l'incertezza) o massimo (entro l'incertezza), dove il massimo, come discusso altrove, si riferisce a un valore legato, nel nostro caso, alla tensione di alimentazione della scheda (tipicamente $V_{ref} \sim 5$ V).

C'è un'interessantissima ulteriore opzione: alcune delle porte digitali di Arduino, quelle marcate sulla scheda con un simbolo tilde, possono operare in modalità *Pulse Width Modulation* (PWM). Questo significa che in uscita si può trovare un'onda quadra con *duty cycle variabile* da zero (onda quadra "spenta", cioè nessun segnale in uscita) al massimo (onda quadra "sempre accesa", cioè segnale continuo pari al massimo, come per le altre porte digitali quando vengono poste a "livello alto"). Il duty cycle è aggiustabile in maniera digitale agendo su un carattere, cioè su un byte: dunque sono possibili $2^8 = 256$ livelli diversi di duty cycle che possono essere scelti via software, attraverso un'opportuna istruzione dello sketch. Più propriamente, questa onda quadra si chiama *treno di impulsi*.

A. Implementazione software

Un aspetto molto importante di Arduino è che il treno di impulsi prodotto è gestito "direttamente" dal microcontroller, cioè esso non risulta da cicli inseriti nel programma dello sketch. Dunque le sue caratteristiche non risentono di eventuali latenze del microcontroller e, una volta definite attraverso l'istruzione relativa, rimangono nominalmente costanti nel tempo (entro l'incertezza).

Come controparte, la frequenza del treno di impulsi è determinata internamente e non può essere variata facilmente. Infatti essa è agganciata alla frequenza del contatore, cioè dell'orologio interno al microcontroller. Inoltre, tale frequenza è, purtroppo, piuttosto bassa, in analogia con la generale "lentezza" del microcontroller usato in Arduino. Infatti essa vale nominalmente $f = 976$ Hz per le porte digitali PWM ~ 5 e ~ 6 , e 488 Hz per le porte digitali PWM ~ 3 , ~ 9 , ~ 10 . Spulciando tra le specifiche, si

vede come in realtà la frequenza possa essere modificata, in particolare aumentata per multipli di 2, ma di questa possibilità, che implica di ritoccare in qualche modo l'orologio interno del microcontroller, non faremo uso.

L'istruzione software da mettere nello sketch per creare un treno di impulsi con un certo duty cycle è molto semplice e auto-esplicativa. Chiamata `RampPin` la variabile che punta al numero della porta digitale PWM da impiegare, che è la ~ 5 nel nostro caso (e infatti nello sketch comparirà la dichiarazione `const unsigned int RampPin = 5;`), si dovrà inizializzare questa porta come uscita attraverso il comando, da mettere nel `void setup` dello sketch, `pinMode(RampPin, OUTPUT);`. A questo punto, un'onda con duty cycle corrispondente al livello i (con i intero compreso tra 0 e 255) sarà ottenuta con il semplice comando `analogWrite(RampPin, i);`.

B. Integratore

Abbiamo dunque capito come creare un treno di impulsi, cioè un'onda quadra con un certo duty cycle. Siamo ancora lontani dall'avere una d.d.p. continua variabile attraverso semplici istruzioni software, però abbiamo disponibile tra le nostre conoscenze un metodo che permette di ottenere una tensione *quasi continua* a partire dal treno di impulsi. È infatti evidente che il *valore medio* di un'onda quadra con duty cycle variabile dipende dal duty cycle stesso. Sappiamo poi che l'operazione di media è (a meno di coefficienti) equivalente all'*integrazione temporale* e ci è ben noto come costruire un integratore, per esempio facendo uso di un filtro passa-basso RC.

Naturalmente il circuito RC che ci proponiamo di realizzare dovrà avere una costante tempo RC sufficientemente alta, ovvero una frequenza di taglio $f_T = 1/(2\pi RC)$ sufficientemente bassa, in modo da permettere un'efficace integrazione temporale. In particolare ci aspettiamo che debba essere $f_T \ll f$; negli esempi ("simulati") riportati nel seguito supporremo di avere $f_T = 10$ Hz, cioè $f/f_T \sim 10^2$.

D'altra parte è ovvio che, aumentando il tempo di integrazione, cioè diminuendo la *banda passante* del sistema, dovremo introdurre degli opportuni tempi di attesa nell'operazione di Arduino, necessari affinché, dopo aver cambiato il duty cycle del treno di impulsi, il segnale in uscita dall'integratore possa raggiungere una nuova condizione stazionaria. Dunque avremo un'acquisizione automatizzata che, però, richiederà un po' di tempo per essere compiuta.

C. Serie di Fourier per il treno di impulsi

È sicuramente interessante "simulare" il comportamento di un integratore al cui ingresso abbiamo un treno di impulsi, ovvero un'onda quadra con un certo duty cycle. Possediamo già lo strumento concettuale che consente di eseguire la simulazione: è sufficiente esprimere il treno

di impulsi in serie di Fourier, cioè conoscerne i coefficienti dell'espansione di Fourier, e quindi applicare alle varie componenti, cioè alle varie armoniche, la funzione di trasferimento [attenuazione $A(f)$ e sfasamento $\Delta\phi$] del passa-basso RC.

Due osservazioni preliminari:

- nella simulazione considereremo una situazione “ideale”, per cui trascureremo le resistenze interne, per esempio quella del generatore (la resistenza della porta digitale di Arduino) e di tutto il resto, cioè del circuito contenente il diodo e il collegamento alle porte analogiche di Arduino necessarie per la misura; vedremo nel seguito che queste ipotesi possono non essere del tutto verificate nell'esperimento;
- il treno di impulsi in ingresso all'integratore non sarà mai alternato: infatti, anche per un'onda quadra simmetrica (duty cycle pari al 50%), Arduino fa sì che la d.d.p. prodotta oscilli tra zero e il valore massimo, cioè sia sempre positiva, per cui la sua media non è mai nulla.

Supponendo per semplicità un treno di impulsi rappresentato da una funzione $g(t)$ *pari* nel tempo (cioè l'istante $t = 0$ si trova a metà strada della parte alta di un impulso) e di ampiezza unitaria (valore minimo 0, valore massimo 1, in unità arbitrarie), si ha che $g(t)$ può essere rappresentata dalla seguente serie di *coseni*:

$$g(t) = \delta + \sum_{k=1}^n \frac{2}{k\pi} \sin(k\pi\delta) \cos(\omega_k t), \quad (2)$$

dove $\delta = \tau/T$ rappresenta il duty cycle variabile tra 0 e 1 (τ è la durata della parte alta dell'impulso e T è il periodo del treno di impulsi) e $\omega_k = k\omega$ è la frequenza angolare dell'armonica k -esima. Poiché talvolta il duty cycle si esprime come percentuale D (la percentuale di tempo in cui l'impulso si trova a livello alto rispetto al periodo), scriviamo l'ovvia conversione $D = \delta \times 100$. Infine, tenendo conto del fatto che in Arduino il duty cycle può essere impostato via software attraverso l'intero `i` variabile tra 0 e 255 (vedi sopra), l'altrettanto ovvia conversione che lega δ a `i` recita $\delta = i/256$.

Come noto, per determinare la funzione $g_{out}(t)$ che rappresenta l'uscita dell'integratore le varie armoniche di frequenza $f = \omega_k/(2\pi)$ vanno moltiplicate per la funzione che esprime il guadagno del passa-basso, $A(f) = 1/\sqrt{1 + (f/f_T)^2}$, e sfasate di $\Delta\phi = \arctan(-f/f_T)$, esattamente come si fece per l'esercizio sulla “pinna di squalo”. La Fig. 2 riporta un esempio dei risultati per vari valori del duty cycle (in legenda si riporta per chiarezza proprio l'istruzione da usare nell'eventuale sketch di Arduino): i grafici mostrano in blu il treno di impulsi simulato e in rosso l'uscita simulata dell'integratore. Come specificato sopra, il treno di impulsi ha una frequenza $f = 976$ Hz, mentre la frequenza di taglio dell'integratore è supposta $f_T = 10$ Hz. Per chiarezza, la Fig. 3 mostra

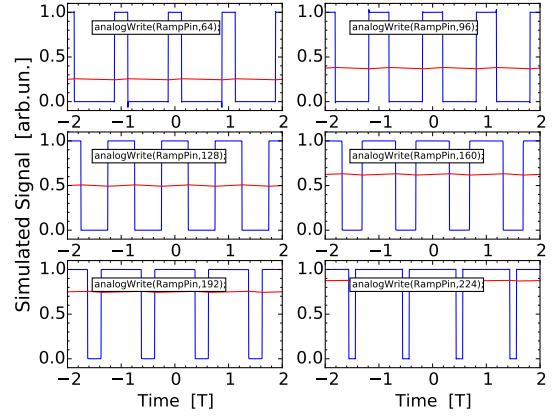


Figura 2. Esempi di simulazione di treni di impulsi con duty cycle variabile (curve blu) e di uscita dall'integratore (curve rosse). I vari grafici si riferiscono a diverse scelte del duty cycle: in legenda è riportata l'istruzione software usata nello sketch di Arduino. Per il calcolo si è supposto un integratore RC con frequenza di taglio $f_T = 10$ Hz; il periodo T con cui si misurano i tempi vale $T = 1/f$, con $f = 976$ Hz.

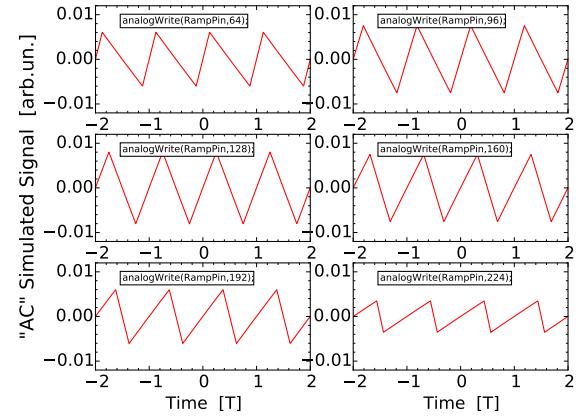


Figura 3. Analogo di Fig. 2 per le sole tracce rappresentative del segnale in uscita dall'integratore: per questa figura, a tale segnale è stato sottratto il valore medio nel tempo, simulando, in pratica, un'osservazione con oscilloscopio “accoppiato in AC”.

il segnale in uscita dall'integratore a cui è stato sottratto il valore medio nel tempo: dunque i pannelli riportati in questa figura rappresentano una sorta di simulazione di quanto si vedrebbe usando un oscilloscopio “accoppiato in AC”, che permette di apprezzare la discrepanza rispetto a un segnale idealmente costante.

Il risultato simulato è in accordo con le attese: effettivamente all'uscita dell'integratore si ritrova un livello *quasi-continuo* che dipende linearmente dal duty cycle impostato. Il carattere quasi-continuo è dovuto alla frequenza di taglio dell'integratore che abbiamo supposto finita, cioè diversa da zero: essa dà luogo a una sorta di

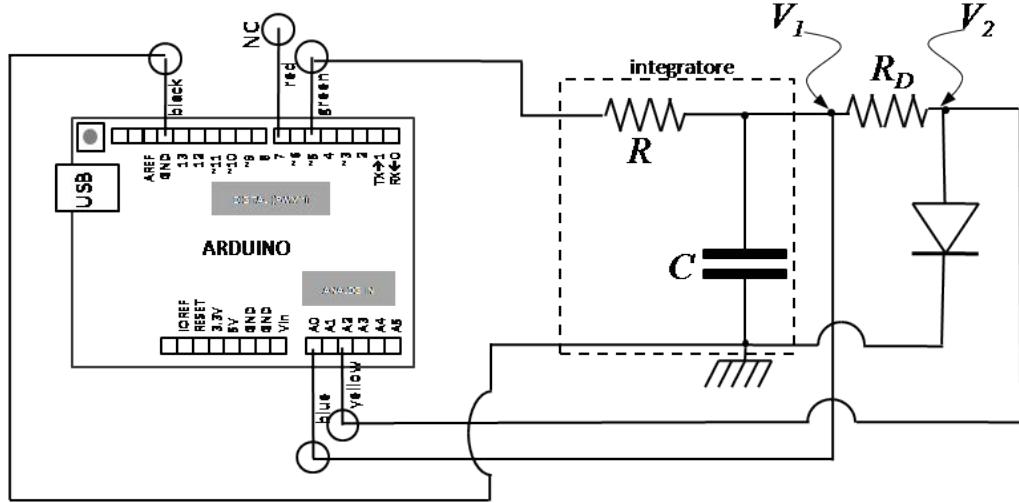


Figura 4. Schema circuitale dell’esperienza con indicate le quattro connessioni da effettuare alla scheda Arduino (NC significa non collegato).

ripple (piccola modulazione) che non riesce a essere integrato in maniera completa e che, come risulta chiaro da Fig. 3, è particolarmente rilevante per i valori intermedi del duty cycle. In linea di principio, l’entità del ripple, che è generalmente “piccola” rispetto al livello quasi-continuo, potrebbe renderne la presenza ininfluente per gli scopi del nostro esperimento: su questo argomento torneremo in seguito.

III. MISURA DELLA CORRENTE

Dobbiamo a questo punto definire il metodo che consente di misurare l’intensità di corrente I che fluisce nel diodo. La soluzione più semplice consiste nell’inserire una resistenza R_D in serie al diodo: infatti, per la legge di Ohm è semplicemente $I = \Delta V_{RD}/R_D$, dove ΔV_{RD} è la caduta di potenziale misurata ai capi di questa resistenza.

Dal punto di vista pratico, dato che le porte analogiche di Arduino misurano d.d.p. relative alla linea di terra, la misura ΔV_{RD} deve essere eseguita per differenza, $\Delta V_{RD} = V_1 - V_2$, tra le d.d.p. (riferite a terra) che si trovano all’ingresso (V₁) e all’uscita (V₂) di R_D . La d.d.p. V₂ è inoltre quella che di fatto si trova ai capi del diodo, cioè corrisponde a ΔV nella simbologia di Eq. 1, per cui la misura indipendente di queste due

d.d.p. consente di trovare tutte le grandezze necessarie per ricostruire la curva I-V.

Oltre alla determinazione di I , la resistenza R_D ha anche un altro ruolo. Quando si trova in conduzione, cioè per ΔV superiore a un valore di soglia V_{thr} (tipicamente compreso nell’intervallo 0.45–0.65 V, per diodi al silicio), il diodo sostiene il passaggio di correnti potenzialmente molto intense [si veda Fig. 1(b)], incompatibili con le possibilità di Arduino. La resistenza in serie al diodo limita la corrente massima richiesta.

Supponendo di porre tale limite sotto alla decina di mA (le porte digitali di Arduino sono progettate per fornire un massimo di 20 mA, secondo datasheet, ma è meglio tenersi al di sotto), e tenendo conto che la d.d.p. erogata è al massimo attorno a 5 V, R_D deve essere dimensionata nell’ordine delle centinaia di ohm, o superiore. Ponendo, ad esempio, $R_D = 680$ ohm, si ha una richiesta di corrente massima inferiore a 8 mA, dove per semplicità si sono trascurate la resistenza interna effettiva del diodo e la resistenza R dell’integratorе.

In definitiva, lo schema circuitale adottato è quello di Fig. 4: in esso si prevede l’impiego della porta digitale PWM ~5 (boccola verde) e delle porte analogiche A0 e A2 (boccole blu e gialla), rispettivamente per la misura delle d.d.p. V₁ e V₂.

IV. SCRIPT DI PYTHON E SKETCH DI ARDUINO

Come nostro solito, ci serviamo di uno script di Python per gestire la partenza delle operazioni di Arduino e per gestire il trasferimento dei dati da questo al computer

tramite porta seriale USB.

Lo script sfrutta tutte le particolarità già discusse a proposito dell’esperienza sulla carica/scarica del condensatore; esso si trova in rete con il nome di `diodo2016.py` e, ovviamente, è già caricato sul computer di laboratorio. Anche in questa esperienza lo script invia un’informazio-

ne ad Arduino attraverso un carattere (byte) che rappresenta il ritardo tra una coppia di misure e la successiva, necessario per permettere all'integratore di raggiungere condizioni stazionarie. Il carattere, che può essere impostato tra 1 e 9, rappresenta il ritardo in unità di 10 ms, e per default è regolato a 50 ms. Poiché nello sketch di Arduino ci sono due istruzioni successive di ritardo, di fatto,

con le impostazioni di default, occorrono $50 \times 2 = 100$ ms per acquisire un singolo punto della curva I-V. Tenendo conto di ulteriori ritardi (necessari per la scarica del condensatore e per evitare impallamenti nella comunicazione seriale), l'acquisizione della curva richiede un tempo stimabile in diverse decine di secondi.

Il testo dello script è il seguente:

```

import serial # libreria per gestione porta seriale (USB)
import time  # libreria per temporizzazione

print('Apertura della porta seriale\n') # scrive sulla console (terminale)
ard=serial.Serial('/dev/ttyACM0',9600) # apre la porta seriale /dev/ttyACM0
time.sleep(2) # aspetta due secondi
ard.write(b'5')#intervallo (ritardo) in unita' di 10 ms <<< questo si puo' cambiare (default 50 ms)
print('Start!\n') # scrive sulla console (terminale)
Directory='../../dati_arduino/' # nome directory dove salvare i file dati
FileName=(Directory+'diodo.txt') # nomina il file dati <<< DA CAMBIARE SECONDO GUSTO
outputFile = open(FileName, "w+" ) # apre file dati in scrittura

# loop lettura dati da seriale (sono 256 righe, eventualmente da aggiustare)
for i in range (0,256):
    data = ard.readline().decode() # legge il dato e lo decodifica
    if data:
        outputFile.write(data) # scrive i dati nel file

outputFile.close() # chiude il file dei dati
ard.close() # chiude la comunicazione seriale con Arduino
print('end') # scrive sulla console (terminale)

```

Anche lo sketch di Arduino, che si può trovare in rete con il nome `diodo2016.ino`, è molto semplice e pressoché auto-esplicativo. Poiché in questo esperimento non è necessario spingere al massimo possibile il rate di campionamento, non vengono incluse le istruzioni necessarie

all'"overclock" del microcontroller (tra l'altro, esse sono incompatibili con le altre operazioni richieste ad Arduino in questa esperienza).

Il testo è il seguente:

```

const unsigned int RampPin = 5; //pin 5 uscita pwm per generare la rampa
const unsigned int analogPin_0=0; //pin A0 per lettura V1
const unsigned int analogPin_2=2; //pin A2 per lettura V2
unsigned int i=0; //variabile che conta gli step durante la salita della rampa
int V1[256]; //array per memorizzare V1 (d.d.p, letta da analogPin_0)
int V2[256]; //array per memorizzare V2 (d.d.p, letta da analogPin_2)
int delay_ms; //variabile che contiene il ritardo tra due step successivi (in unita' di 10 ms)
int start=0; //flag per dare inizio alla misura

//Inizializzazione
void setup()
{
    pinMode(RampPin, OUTPUT); //pin pwm RampPin configurato come uscita
    Serial.begin(9600); //inizializzazione della porta seriale
    Serial.flush(); // svuota il buffer della porta seriale
}

//Ciclo di istruzioni del programma
void loop()

```

```

{
  if (Serial.available() >0) // Controlla se il buffer seriale ha qualcosa
  {
    delay_ms = (Serial.read()-'0')*10; // Legge il byte e lo interpreta come ritardo (unita' 10 ms)
    Serial.flush(); // Svuota il buffer della seriale
  start=1; // Pone il flag start a uno
  }
  if(!start) return // solo se il flag Ã“ a uno parte l'acquisizione
  delay(500); // attende 0.5s per evitare casini
  if (start==1)
    analogWrite(RampPin,0); // all'inizio pone a 0 la RampPin per favorire scarica condensatore
  delay(1500); // attende 1.5s per scaricare condensatore
  for(i=0;i<256;i++) //il valore che definisce il duty cycle dell'onda quadra e' scrivibile su 8 bit
    //cioe' assume valori da 0-->duty cycle 0% a 256-->duty cycle 100%
  {
    analogWrite(RampPin, i); //incrementa il duty cycle di uno step
    delay(delay_ms); //aspetta il tempo impostato
    V1[i]=analogRead(analogPin_uno); //legge il pin analogPin_uno
    V2[i]=analogRead(analogPin_due); //legge il pin analogPin_due
    delay(delay_ms); //aspetta il tempo impostato
  }
  for(i=0;i<256;i++) //nuovo ciclo che scorre gli array di dati e li scrive sulla seriale
  {
    Serial.print(V1[i]);
    Serial.print(" ");
    Serial.println(V2[i]);
  }
  start=0; // Annulla il flag
  Serial.flush(); // svuota il buffer della porta seriale
}

```

V. CALIBRAZIONE E INCERTEZZE

Lo scopo dell'esperienza è quello di costruire una curva I-V da cui, oltre all'andamento, sia anche possibile dedurre i valori caratteristici in gioco, espressi nelle debite unità fisiche. Pertanto è opportuno che le grandezze digitalizzate ([digit]) siano convertite in unità [V].

A questo scopo è possibile impiegare la procedura di calibrazione “alternativa”, che consiste nella misura di V_{ref} e nella determinazione del fattore di calibrazione $\xi = V_{ref}/1023$ (le sue unità di misura sono [V/digit], o, ancora meglio, [mV/digit]). Come già discusso in altra sede, V_{ref} è atteso coincidere, almeno approssimativamente, con il massimo valore della tensione fornita dalle uscite digitali di Arduino. Poiché al termine dello sketch la porta ~ 5 si trova al livello “alto” (e ci rimane finché l'acquisizione non viene fatta ripartire), per conoscere V_{ref} è sufficiente misurare con il multmetro digitale, al termine delle acquisizioni, la d.d.p. fra tale porta digitale e la linea di massa, o terra, cioè tra la boccola verde e quella nera di Arduino.

È ovvio che questa operazione di misura deve essere condotta a circuito aperto, cioè scollegando quanto si trova di seguito alla porta (almeno il diodo). Infatti, come già affermato, nel diodo può circolare corrente di inten-

sità tutt'altro che trascurabile, che quindi provoca una caduta di tensione sulla serie di resistori R e R_D . Nell'eseguire l'esperimento può essere istruttivo verificare qual è la differenza nella lettura di V_2 quando il diodo viene collegato o scollegato (si ricorda, en passant, che scollegare un componente a due fili significa scollegare uno o tutti e due i fili, e basta). Dunque dalla singola misura è possibile ricavare il fattore di calibrazione e l'incertezza ad esso associata, dovuta all'errore (calibrazione e lettura, sommate in quadratura) della misura di V_{ref} con il multmetro.

Come sappiamo da precedenti esperienze, la calibrazione “alternativa” trascura completamente la non linearità del digitalizzatore, in particolare l'eventuale presenza di un offset. Per questa esperienza, e a meno di non eseguire una calibrazione “completa” della scheda Arduino (che richiede tempo e sforzi), possiamo disinteressarci di questo aspetto, tenendo conto che l'Eq. 1, che useremo per interpretare i dati, contiene già, di fatto, un termine costante. Pertanto non avrebbe molto senso aggiungere un ulteriore offset all'equazione stessa, con lo scopo di migliorare l'accordo tra dati e best-fit.

Per determinare l'incertezza delle misure possiamo usare il consueto approccio per cui attribuiamo un errore convenzionale (di lettura) di ± 1 digit alla lettura digi-

talizzata. Poiché in questa esperienza convertiamo la lettura digitalizzata in unità fisiche, dovremo sommare (in quadratura) a questo errore quello dovuto alla calibrazione di Arduino. Questa operazione è sufficiente a determinare l'incertezza sulla grandezza che sta sull'asse orizzontale del grafico che intendiamo costruire, cioè $\Delta V = V_2$ (chiameremo il suo errore δV).

Invece la grandezza che compare sull'asse verticale del grafico che vogliamo produrre, cioè l'intensità di corrente I , è valutata attraverso la relazione

$$I = \frac{V_1 - V_2}{R_D}, \quad (3)$$

dove tutte le grandezze sono già state definite in precedenza. Per stimarne l'incertezza δI dovremo fare debito uso delle regole di propagazione dell'errore. Poiché al numeratore di Eq. 3 compare una differenza tra grandezze affette dalla stessa incertezza di calibrazione, dovremo porre attenzione a non sovrastimare l'errore della differenza (questo argomento dovrebbe essere nelle vostre conoscenze di analisi dati).

VI. FUNZIONAMENTO EFFETTIVO DELL'INTEGRATORE

Ora che abbiamo specificato più nel dettaglio come intendiamo effettuare le misure, possiamo tornare all'analisi del funzionamento dell'integratore. In buona sostanza, esso si comporta come il *livellatore* di un alimentatore basato su trasformatore e raddrizzatore a semionda. Nell'esaminare questa tipologia di circuiti si ottiene che il ripple può dipendere dal "carico": in particolare, esso tende ad aumentare quando il carico (supposto resistivo) diminuisce, cioè quando viene richiesta una maggiore corrente.

Nell'analisi di Sez. II B abbiamo di fatto *trascurato* il carico, e tutti gli eventuali problemi connessi: abbiamo infatti supposto di trascurare tutte le resistenze al di fuori di quella inserita nell'integratore. Invece, nella realtà del nostro circuito al "livellatore", cioè al condensatore C , viene richiesta corrente per tutta il tempo che intercorre tra due impulsi successivi del treno. La corrente fluisce attraverso R_D e quindi nel diodo, quando la d.d.p. ai suoi capi è tale da portarlo in conduzione; ovviamente, in seguito a questo processo il condensatore perde parte della carica che ha accumulato in precedenza, quando l'impulso si trovava a livello "alto", per cui il meccanismo di livellamento tende a perdere di efficacia.

L'effetto del carico potrebbe essere analizzato sulla base di un opportuno modello (questo potrebbe essere tentato, ad esempio, in una relazione semestrale). Qui ci limitiamo a notare che, nelle condizioni effettive di funzionamento, il ripple della d.d.p. applicata al diodo può rivelarsi sensibilmente superiore rispetto a quanto predetto in Fig. 3: siete invitati a verificare sperimentalmente la sua entità, per esempio monitorando con l'oscilloscopio il segnale V_2 durante la misura.

La presenza del ripple, combinata con la circostanza che le digitalizzazioni di V_1 e V_2 non sono simultanee fra loro (fra di esse intercorre un tempo almeno pari a quello necessario per la singola digitalizzazione), può produrre delle "fluttuazioni" nella misura di V_2 e di I . È evidente che tale problema potrebbe essere mitigato aumentando R_D . Tuttavia, così facendo si determinerebbe una rilevante caduta di potenziale su questo componente, che finirebbe per limitare il range di valori di ΔV esplorati (tutto questo diverrà più chiaro quando saranno presentati i risultati esempio). D'altro canto, come già sotto-lineato, questo componente serve anche per limitare la richiesta di corrente alle porte di Arduino, per cui R_D non può nemmeno essere scelta troppo bassa (un limite minimo ragionevole può essere il centinaio di ohm).

Una possibilità alternativa consiste nello "spianare" al meglio il treno di impulsi, cioè scegliere una frequenza di taglio f_T particolarmente bassa, ovvero, rifrasando, usare R e C di valore alto. Poiché R è di fatto in serie a R_D e al diodo, conviene che anche questa resistenza non abbia valore troppo alto (si consiglia R dell'ordine delle poche centinaia di ohm). Quindi l'opzione più opportuna può essere quella di impiegare un condensatore C di capacità elevata.

Per mantenere dimensioni fisiche accettabili in condensatori di elevata capacità, è in genere necessario servirsi di componenti che sfruttano tecnologie diverse rispetto a quella dei dispositivi a poliestere (detti anche "a carta") che siamo soliti usare. In particolare, nell'esperienza potrebbe essere consigliabile servirsi di *condensatori elettrolitici*, che, a parità di dimensioni fisiche, consentono capacità nettamente superiori.

A causa delle loro caratteristiche costruttive (che siete fortemente invitati a studiare), i condensatori elettrolitici sono componenti a due fili *polarizzati*, cioè il loro collegamento deve essere effettuato solo in un senso. Il "polo" negativo del componente, normalmente collegato alla carcassa metallica del suo involucro esterno, deve essere collegato alla linea di massa, o terra.

VII. ESEMPIO DI MISURA E COMMENTI

L'esperimento esempio qui riportato è stato compiuto usando un integratore realizzato con $R = 330$ ohm (nominali, tolleranza 5%) e $C = 100 \mu\text{F}$ (nominali, tolleranza 20%), capacità ottenuta proprio con un condensatore elettrolitico. La frequenza di taglio nominale è quindi $f_T = 4.8$ Hz, con tolleranza dominata da quella sulla capacità. La resistenza in serie al diodo è stata misurata, ottenendo $R_D = (682 \pm 6)$ ohm. La tensione di riferimento, misurata a circuito aperto con la procedura descritta in precedenza, è risultata $V_{ref} = (5.02 \pm 0.03)$ V. Il fattore di calibrazione è quindi $\xi = (4.91 \pm 0.03)$ mV/digit. La somma delle resistenze $R + R_D$ è relativamente grande, oltre 1 kohm, e quindi la richiesta di corrente alla porta di Arduino è relativamente piccola, limitata nominalmente a circa 5 mA. Come intervallo di tempo tra un'acquisizio-

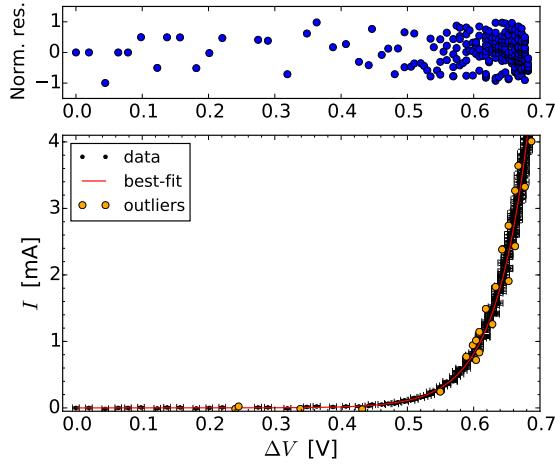


Figura 5. Esempio di misure ottenute con la scelta dei parametri e dei componenti circuituali indicati nel testo; la linea rossa continua rappresenta il best-fit condotto secondo l’Eq. 1 (commenti e risultati nel testo), i marker arancioni individuano i potenziali outliers, secondo quanto specificato nel testo. Il pannello superiore riporta il grafico dei residui normalizzati (depurati dagli outliers).

ne e la successiva è stato mantenuto il valore di default di 50 ms (la scelta non è critica).

La Fig. 5 riporta un esempio delle misure: nella determinazione delle barre di errore sono state seguite le istruzioni specificate in precedenza. La figura mostra anche, con una linea continua rossa, il risultato di un best-fit dei dati secondo l’Eq. 1, eseguito lasciando come parametri liberi del fit I_0 e ηV_T , il corrispondente grafico dei residui normalizzati (pannello superiore), e i potenziali *outliers*, identificati come sarà discusso nel seguito e indicati con pallini arancioni sovrapposti al grafico dei dati sperimentali.

Facciamo subito alcune osservazioni preliminari sui dati: in primo luogo l’andamento qualitativo è quello atteso, cioè, grazie anche alla scala lineare di rappresentazione (potrebbe essere opportuno impiegare una rappresentazione semi-logaritmica, che però darebbe un risultato visualmente meno immediato), si intuisce una crescita pressoché esponenziale della corrente I in funzione di ΔV , in particolare per $\Delta V \gtrsim 0.5$ V.

Accanto a questa osservazione in linea con le aspettative, ci sono alcuni aspetti apparentemente meno ovvi. Per esempio, si nota che la spaziatura dei valori sull’asse orizzontale, che in principio ci aspetteremmo sempre uguale (approssimativamente pari $V_{ref}/256 \sim 20$ mV, essendo 256 i livelli su cui viene aggiustato il duty cycle del treno di impulsi), non è affatto uniforme. Per bassi valori di ΔV possiamo attribuire la circostanza all’incertezza con la quale viene generata la d.d.p. e al tempo di risposta finito dell’integratore, ma certamente l’addensamento dei punti di misura a partire da $\Delta V \sim 0.5$ V richiede una spiegazione. Inoltre è anche evidente che il range spaz-

zato nella misura è molto inferiore rispetto alle aspettative: infatti il ΔV massimo esplorato non raggiunge 0.7 V, mentre invece ci aspetteremmo, sulla base di un ragionamento molto naïf, valori prossimi a V_{ref} .

Il motivo può essere facilmente intuito ricordando le caratteristiche di funzionamento del diodo bipolare a giunzione e osservando come, per $\Delta V \gtrsim 0.6$ V, l’intensità di corrente I salga a valori dell’ordine di alcuni mA. In queste condizioni la caduta di potenziale sulla serie $R + R_D$ è tutt’altro che trascurabile (essa vale oltre 4 V per $I = 4$ mA), per cui la d.d.p. ΔV effettivamente applicata al diodo si riduce considerevolmente. Rifrasando, potremmo affermare che il generatore che alimenta il diodo ha una resistenza interna (“a la Thévenin”) tutt’altro che trascurabile, stimabile proprio come la somma $R + R_D$ (se si suppone praticamente nulla la resistenza interna della porta digitale di Arduino).

In buona sostanza, come evidenziato anche dall’andamento della curva caratteristica I-V, vedi Fig. 1(b), attorno o sopra il valore di soglia V_{thr} (ricordiamo che esso è tipicamente compreso tra 0.45 e 0.65 V, per un ordinario diodo al silicio), la d.d.p. ΔV tende proprio a stabilizzarsi attorno a tale valore, indipendentemente (o “poco dipendentemente”) dalla corrente. Questa affermazione è naturalmente da prendere con le molle, se non altro perché V_{thr} non è definito dal punto di vista matematico. La definizione pratica che ne è talvolta data, come la d.d.p. applicata al diodo tale che in esso fluisca una corrente di intensità pari a 1/100 dell’intensità massima tollerata (300 mA, supponendo che il diodo impiegato sia il modello 1N914 al cui datasheet si fa riferimento), conduce a $V_{thr,prat} \simeq 0.66$ V, dove il pedice *prat* indica che si tratta di un valore riferito a una definizione pratica. Esso è leggermente al di fuori del range generalmente accettato, e la (piccola) discrepanza può essere interpretata come dovuta alla caduta di potenziale sulle resistenze che si trovano in serie alla giunzione, cioè i pezzi di semiconduttore drogato p e n, gli elettrodi, i conduttori di collegamento, che non sono considerate quando si descrive il modello di un diodo a giunzione “ideale”; la presenza di queste resistenze non è ovviamente considerata nell’equazione modello (Eq. 1). Notate in ogni caso che esistono anche altre definizioni pratiche di soglia, che possono condurre a valori leggermente diversi.

Allora, l’infittirsi dei punti sperimentali rispecchia la circostanza che, a partire da un certo valore del duty cycle, V_2 , cioè ΔV in Fig. 5, tende ad aumentare di pochissimo da una misura alla seguente. Poiché, inoltre, I è costruita a partire dalla differenza $V_1 - V_2$, anche a causa del ripple residuo in uscita dall’integratore può verificarsi che i dati acquisiti non siano ordinati monotonicamente. Tutto questo dà luogo a una sorta di “fluttuazioni” dei dati graficati.

Chiariti questi aspetti, torniamo ad occuparci del best-fit. Esso è stato condotto tenendo in conto anche l’incertezza δV sulla grandezza riportata in asse orizzontale, che l’estensione delle barre di errore suggerisce non trascurabile. Allo scopo si è determinato l’“errore equivalente” at-

traverso propagazione dell'errore applicata all'equazione modello (Eq. 1), che è poi stato sommato in quadratura con δI (a sua volta determinato con attenzione, secondo quanto specificato in precedenza). Naturalmente la procedura ha richiesto inizialmente di individuare i parametri del best-fit (I_0 e ηV_T) usando solo δI , esattamente come abbiamo già fatto in altre circostanze e come tutti ben sapete da lungo tempo.

I risultati del best-fit, per il quale è stata usata l'opzione `absolute_sigma = False` (le incertezze hanno sicuramente un carattere non prevalentemente statistico, vista la presenza di errori sistematici di calibrazione), sono

$$I_0 = (5.2 \pm 0.2) \text{ nA} \quad (4)$$

$$\eta V_T = (50.1 \pm 0.3) \text{ mV} \quad (5)$$

$$\chi^2/\text{ndof} = 102/254 \quad (6)$$

$$\text{Norm. cov.} = 0.997 . \quad (7)$$

L'elevato valore della covarianza normalizzata è conseguenza del modello, Eq. 1: in esso, se si considerano punti sperimentali tali che il termine esponenziale prevalga sull'unità (praticamente tutti quelli acquisiti per ΔV superiore a poche decine di mV), c'è una correlazione pressoché totale tra I_0 , che sta a moltiplicare l'esponenziale, e ηV_T , che sta a dividere nell'argomento dell'esponenziale.

In termini assoluti, i valori trovati per I_0 e per ηV_T sono grossolanamente compatibili con le aspettative. In particolare, facendo riferimento al datasheet del diodo 1N914 (un "paradigma" dei diodi al silicio di piccola potenza), si osserva come valori tipici per I_0 siano compresi tra circa 4 e circa 6 nA (la misura del datasheet è in realtà riferita a una polarizzazione inversa $\Delta V = -1$ V) e, in ogni caso, correnti di saturazione inversa di alcuni nA, e fino a oltre 10 nA, sono ampiamente ragionevoli. Tenendo poi conto che η , per un ordinario diodo al silicio, è generalmente compreso tra 1.5 e 2, anche il valore di

ηV_T è perfettamente in linea con le aspettative, considerando che la misura è eseguita a temperatura ambiente, almeno finché si considera trascurabile il riscaldamento dovuto a effetto Joule del diodo (un'evoluzione dell'esperimento potrebbe comprendere misure condotte a diverse temperature).

Infine, nonostante i dati presentati in questo esempio non richiedano particolari "attenzioni" di tipo cosmetico, è stata applicata una procedura volta ad identificare eventuali *outliers*, che qui definiamo, in accordo con una precedente esperienza, come dati che distano dalla previsione del best-fit per oltre una certa soglia. Questa procedura, che ha sempre un carattere *arbitrario*, può essere utile per isolare dati affetti da errori di tipo prevalentemente sistematico, legati ad aspetti specifici del funzionamento di Arduino o dell'esperimento. Come in precedenti esperienze, potrebbe infatti verificarsi che la digitalizzazione di segnali variabili nel tempo avvenga in maniera imperfetta; inoltre le "fluttuazioni" dei dati, le cui origini sono state accennate in precedenza, e la possibilità di registrare segnali spuri, legati ad esempio a spikes (impulsi) che circolano all'interno del microcontroller per cause o stocastiche, o dovute a qualche operazione transiente a carico del microcontroller stesso (accensione o spegnimento di porte, o altro), possono anche dare luogo ad artefatti potenzialmente rilevanti.

In questo esempio, è stato arbitrariamente scelto di considerare potenziali outliers i dati che si discostavano per oltre una barra di errore (comprensivo dell'"errore equivalente") dalle previsioni del best-fit: i dati così identificati, in numero totale di 23, sono marcati con un pallino arancione in Fig. 5. Come si può facilmente intuire, in questo esempio la rimozione degli outlier dal set di dati considerato nel best-fit non produce modifiche sostanziali ai risultati riportati in precedenza, a parte un'ovvia diminuzione del χ^2 (che scende a 52 per $\text{ndof} = 231$).

Esercizi obbligatori su serie di Fourier e Python

francesco.fuso@unipi.it

(Dated: version 8 - FF, 29 novembre 2019)

Questa nota propone alcuni esercizi sullo sviluppo in serie di Fourier che dovete svolgere, da soli o in piccolo gruppo, con Python. Prima del testo degli esercizi, si dà una rapida occhiata al background matematico e alle definizioni necessarie. Inoltre questa nota riporta anche alcuni hints per l'esecuzione degli esercizi e il risultato trovato da me, che può essere utile come referenza.

I. SVILUPPO IN SERIE DI FOURIER

Come si dimostra (in matematica), una *qualsiasi* funzione periodica $g(t)$, con periodo $T = 1/f = 2\pi/\omega$, può essere espressa come sovrapposizione lineare di una costante e di funzioni armoniche di pulsazione ω e $\omega_k = k\omega$, con k intero positivo, dette rispettivamente armonica fondamentale e armoniche superiori di ordine k . Esistono vari modi per esprimere questa sovrapposizione, che dipendono fondamentalmente dall'impiego di grandezze reali o complesse e dall'uso di termini di fase costanti. Per i nostri scopi attuali, il modo migliore consiste nell'usare una sovrapposizione (serie) di funzioni seno e coseno, da cui il nome di *serie di Fourier di seni e coseni*:

$$g(t) = \frac{a_0}{2} + \sum_{k=1}^n b_k \cos(\omega_k t) + \sum_{k=1}^n c_k \sin(\omega_k t) , \quad (1)$$

dove $n \rightarrow \infty$ e il coefficiente a_0 tiene conto del valore medio della funzione. Quando possibile, nel seguito faremo riferimento a funzioni periodiche *alternate*, per le quali $a_0 = 0$. I *coefficients* dell'espansione di Fourier in seni e coseni, b_k e c_k , si ricavano dalle espressioni

$$b_k = \frac{2}{T} \int_0^T g(t) \cos(\omega_k t) dt \quad (2)$$

$$c_k = \frac{2}{T} \int_0^T g(t) \sin(\omega_k t) dt . \quad (3)$$

Non è di certo questa la sede per entrare nel merito della tanta matematica che è coinvolta nell'argomento (definizioni, dimostrazioni, applicazioni, alcune strepitosamente importanti); mi piace però sottolineare che, in qualche modo, l'approccio permette di sviluppare su una base ortogonale una funzione *periodica* *qualsiasi*. Esso può essere considerato come caso limite di un approccio più generale (“analisi di Fourier”) che consente di scrivere in forma integrale funzioni dipendenti in *qualsiasi* modo dal tempo basandosi ancora su una base fatta di funzioni armoniche. Osservate che il metodo che intendiamo sviluppare trasforma le informazioni ottenute “in un dominio” (per esempio, nel dominio dei tempi, cioè analizzando il comportamento di un sistema in funzione del tempo) in informazioni rilevanti per il dominio “coniugato”, o reciproco (per esempio, nel dominio delle frequenze, dove spesso le informazioni hanno la forma di “spettri”). Questa possibilità ha ricadute bellissime, come avrete modo di verificare nel futuro in altri corsi

e anche, in forma limitata agli aspetti più squisitamente legati all'analisi dei dati, in questo nostro corso.

Per i nostri scopi attuali è subito evidente che l'approccio della serie di Fourier ha conseguenze notevolissime. Infatti abbiamo già sviluppato una tecnica, il metodo simbolico (quello dei “fasori”, con linguaggio da elettrotecnici), che consente di determinare in modo semplice e efficace la risposta di un circuito, ad esempio un filtro, in regime sinusoidale. Poder scomporre un *qualsiasi* segnale periodico in armoniche consente di applicare a *ognuna* di queste la *funzione di trasferimento* del circuito. Ri-sommando le armoniche, cioè usando il principio di sovrapposizione, è infine possibile determinare la risposta del circuito quando ad esso è applicata una forma d'onda periodica *qualsiasi*, che per noi vuol dire, al momento, quadra o triangolare. Vedremo le potenzialità dell'approccio negli esercizi 2-4.

A. Sviluppo di onda quadra e onda triangolare

I coefficienti di Fourier dipendono ovviamente dalla forma della funzione $g(t)$ che si sta considerando. Essi possono essere calcolati attraverso le relazioni Eqs. 2, 3, che richiedono di svolgere integrali sul tempo. Normalmente il calcolo di integrali si esegue, con grande efficacia, con metodi numerici e nell'ambito dell'analisi di Fourier sono stati sviluppati degli algoritmi potentissimi, chiamati in genere *FFT* (Fast Fourier Transform), o *DFT* (Discrete Fourier Transform), che permettono di calcolare numericamente i coefficienti anche per funzioni non periodiche nel tempo. Su tutto ciò torneremo più avanti nel corso.

Per ora, il nostro obiettivo è quello di determinare *analiticamente* i coefficienti dell'espansione di Fourier di semplici funzioni periodiche, precisamente di funzioni tipo onda quadra e onda triangolare (simmetriche e alternate).

Prendiamo come $g(t)$ un'onda quadra di ampiezza $1/2$ (in unità arbitrarie), dunque di ampiezza picco-picco unitaria in modo da semplificare la matematica: potremo sempre riscalarne attraverso semplice moltiplicazione il valore dei coefficienti per aggiustare l'ampiezza. Sempre per comodità, facciamo in modo che l'onda quadra sia alternata (come già affermato, $a_0 = 0$) e dispari; questo vuol dire che essa vale $-1/2$ per $(-T/2, 0)$ e $+1/2$ per $(0, T/2)$. In altre parole scegliamo per comodità l'origine dei tempi in modo tale che l'onda esca fuori proprio in

questo modo. Anche in questo caso non si perde in generalità, dato che, se l'onda quadra considerata non fosse dispari, potremmo sempre immaginare di aggiungere un termine di fase costante per renderla dispari. Rifrasando: i risultati che seguono dipendono dalle scelte fatte, ma l'estensione a casi generali è immediata.

Avendo scelto la nostra onda quadra dispari, è ovvio che ci aspettiamo che il suo sviluppo in serie di Fourier non contenga le armoniche espresse da coseni, cioè possiamo subito porre $b_k = 0$ (per ogni k intero positivo).

Concentriamoci quindi sui coefficienti c_k , iniziando per esempio con il calcolo esplicito di c_1 ($k = 1$). Allo scopo, dobbiamo calcolare l'integrale del prodotto tra la nostra $g(t)$ e la funzione $\sin(\omega t)$ nel “primo” periodo, cioè tra gli estremi di integrazione $(0, T)$. Il pannello in alto di Fig. 1 mostra queste due funzioni e il loro prodotto: l'integrale da calcolare è rappresentato dall'area (segnata) colorata in rosa, che è evidentemente non nulla. Si vede che l'integrale da calcolare è infatti

$$c_1 = 2 \frac{2}{T} \int_0^{T/2} \frac{\sin(\omega t)}{2} dt = \quad (4)$$

$$= \frac{2}{\omega T} \int_0^{\omega T/2} \sin \xi d\xi = \quad (5)$$

$$= -\frac{2}{2\pi} \cos \xi|_0^\pi = \quad (6)$$

$$= \frac{2}{\pi}, \quad (7)$$

dove la vergognosa prolissità dei passaggi dovrebbe fugare ogni dubbio.

Il pannello centrale di Fig. 1 mostra le stesse funzioni per $k = 2$: si vede subito che l'area (segnata) è stavolta nulla, per cui $c_2 = 0$. Il pannello in basso della stessa figura si riferisce invece a $k = 3$, dove l'area non è nulla, per cui $c_3 \neq 0$. Stavolta l'integrale da calcolare è

$$c_1 = 2 \frac{2}{T} \int_0^{T/6} \frac{\sin(3\omega t)}{2} dt = \quad (8)$$

$$= \frac{2}{3\omega T} \int_0^{3\omega T/6} \sin \xi d\xi = \quad (9)$$

$$= -\frac{2}{2\pi} \frac{1}{3} \cos \xi|_0^\pi = \quad (10)$$

$$= \frac{2}{3\pi}. \quad (11)$$

Andando avanti con le armoniche, ci si rende facilmente conto che $c_k = 0$ per tutti i k pari, mentre per i termini dispari si ha

$$c_k = \frac{2}{k\pi}, k \text{ dispari}. \quad (12)$$

Si può dimostrare con un po' di lavoro che, anche rilassando la condizione sulla disparità della funzione, lo sviluppo di un'onda quadra contiene solo armoniche dispari.

Per l'onda triangolare, che per semplicità supponiamo pari, a media nulla, sempre di ampiezza $1/2$ e ampiezza picco-picco unitaria, i conti sono un po' più complicati.

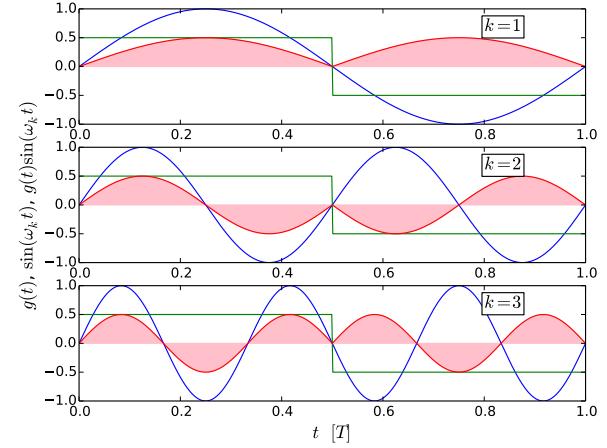


Figura 1. Figura esemplificativa del calcolo dei primi tre coefficienti di Fourier c_k ($k = 1-3$) per un'onda quadra dispari alternata di ampiezza picco-picco unitaria. La scala orizzontale si estende su un singolo periodo. Le linee verdi rappresentano l'onda quadra, le blu le funzioni $\sin(\omega_k t)$, le rosse il prodotto tra le due. In rosa è ombreggiata l'area sottesa al prodotto tra le funzioni, che rappresenta graficamente l'integrale che deve essere calcolato.

Per cominciare, avendo scelto un'onda pari è ovvio aspettarsi che il suo sviluppo in serie di Fourier non contenga le armoniche espresse da seni, cioè possiamo subito porre $c_k = 0$. Come ulteriore passo osserviamo la Fig. 2, concettualmente analoga alla Fig. 1: si vede subito che anche in questo caso l'area sottesa al prodotto tra le funzioni è nulla per $k = 2$, e, per estensione, per tutti i k pari. Dunque anche stavolta lo sviluppo contiene solo i termini con k dispari. Il calcolo, però, non è banale già a partire dal primo termine non nullo, b_1 . Potete provare a operare “per parti”, giungendo a $b_1 = 4/\pi^2$.

Iterando la procedura, cioè andando avanti con le armoniche, si trova

$$b_k = \left(\frac{2}{k\pi} \right)^2, k \text{ dispari}. \quad (13)$$

Un modo più semplice per giungere allo stesso risultato consiste nel notare che un'onda triangolare (pari) può essere considerata come l'integrale nel tempo di un'onda quadra (dispari). Integrare nel tempo le armoniche di tipo seno che compaiono nello sviluppo dell'onda quadra significa trasformarle, segni a parte, in armoniche di tipo coseno, per cui, se per l'onda quadra sono nulli i coefficienti b_k con k dispari, per la triangolare lo sono i c_k , sempre con k dispari. Inoltre quando si integrano termini del tipo $\sin(\omega_k t)$ l'operazione di integrazione fa comparire, messo in evidenza, un termine $1/\omega_k = 1/(k\omega)$, che è responsabile per lo specifico valore dei coefficienti b_k di Eq. 13: essi sono in sostanza il quadrato dei coefficienti c_k di Eq. 12.

Ricapitolando, nel caso di forme d'onda sia quadre che triangolari lo sviluppo in serie di Fourier contiene solo

armoniche dispari. Supponendo le condizioni di parità impiegate, le funzioni che compaiono nello sviluppo sono solo coseni o seni. Inoltre i coefficienti dello sviluppo dipendono fortemente dal tipo di forma d'onda, essendo quelli per l'onda triangolare il quadrato di quelli per l'onda quadra. Rifrasando e tenendo conto della dipendenza dei coefficienti dall'ordine dell'armonica k , potremo concludere che il contenuto armonico delle onde triangolari è diverso da quello delle onde quadre perché in queste ultime i coefficienti scemano più lentamente con l'ordine delle armoniche, ovvero un'onda quadra è “più ricca” di armoniche ad alta frequenza. In qualche modo questo risultato ha a che vedere con la circostanza che due strumenti musicali a corda come il pianoforte e il violino producono suoni parecchio diversi fra di loro.

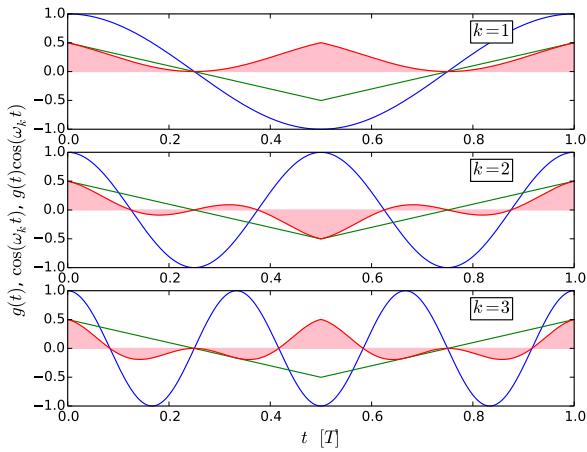


Figura 2. Analogo di Fig. 1 per il caso di onda triangolare alternata pari e ampiezza picco-picco unitaria.

II. ESERCIZIO E HINTS

L'esercizio obbligatorio, da svolgere individualmente o in piccolo gruppo e da consegnare *tassativamente* entro la data stabilita, è in realtà costituito da diversi sotto-esercizi. Per la consegna, è sufficiente che carichiate sulla pagina di e-learning un *unico* file, possibilmente pdf, contenente l'esito (i grafici) e gli script da voi prodotti via e-mail, assieme, se volete, a una semplice e concisa descrizione.

Lo scopo generale dell'esercizio è quello di “divertirsi” (!) costruendo espansioni di Fourier a partire da un foglio bianco, cioè senza usare pacchetti o comandi prefabbricati, e nel modo più semplice possibile (gli script devono essere brevi!). La finalità di questo divertimento è quella di replicare, se preferite “simulare” (ma il termine qui è un po' troppo pomposo), forme d'onda che avete visto in laboratorio, cioè onde quadre e triangolari e la “pinna di squalo” che si osserva, per certe scelte dei parametri di funzionamento, con un integratore RC alimentato da

un'onda quadra. In questo caso, grazie all'acquisizione con Arduino dei dati sperimentali, avete addirittura la possibilità di sovrapporre in un unico grafico esperimento e simulazione, in una sorta di metafora semplice semplice di quanto si fa normalmente in fisica.

Prima di procedere oltre può essere utile soffermarsi su alcuni banali consigli generali per l'implementazione pratica degli script di Python.

In primo luogo è ovvio che si ha a che fare con arrays unidimensionali, cioè vettori: la variabile indipendente (il tempo) è un array, ogni armonica è un array, la funzione ottenuta sommando le armoniche è un array. Lavorare con gli arrays significa fare uso del pacchetto `numpy`. Per esempio, per creare un array (qui chiamato `t`) per la variabile indipendente `t` si può usare il comando `t = numpy.linspace(-2, 2, 5000)`, che genera un vettore di 5000 punti (dovrebbero essere sufficienti per evitare *aliasing* e problemi di sotto-campionamento, ma fate sempre attenzione) distribuiti in modo equispaziato tra -2 e 2. Un comando per creare un array nullo, sempre di 5000 punti, è `w=numpy.zeros(5000)` (`w` è il nome dell'array), se vogliamo invece un array che contenga il coseno della variabile `t` occorre scrivere `w=numpy.sin(t)`, e così via.

Quindi ognuna delle armoniche può essere definita calcolando l'appropriata funzione, per esempio `wk = numpy.cos(omegak*t)` crea una funzione coseno dell'argomento specificato. Infine le varie armoniche, debitamente moltiplicate per i propri pesi (i coefficienti di Fourier), possono essere facilmente sommate tra loro come arrays. Si possono adottare diverse tecniche per istruire il software a eseguire le somme necessarie. Probabilmente il più semplice consiste nell'introdurre dei *cicli* nello script. Questo può per esempio essere fatto con l'istruzione `for counter in range (start, stop, step)`: (ricordate che le righe di script che appartengono al ciclo *devono essere indentate*, cioè scritte premettendo una tabulazione, e fate attenzione ai due punti a fine istruzione), dove la possibilità di introdurre lo step size può essere utile per considerare solo armoniche pari o dispari (è sufficiente porre lo step pari a 2 e regolare opportunamente lo start, per esempio far partire la somma da 1 o da 2).

Naturalmente la somma di Eq. 1 dovrebbe, in matematica, comprendere infiniti elementi. Dal punto di vista numerico la somma sarà estesa a un numero finito, *sufficientemente grande*, di elementi. In genere, per capire se il numero di iterazioni nella somma è sufficientemente grande basta guardare il grafico che si ottiene e verificare “a occhio” se, con la risoluzione adottata (numero di punti dell'array e numero di punti della rappresentazione), la forma è quella attesa, in particolare che non compaiano “spigolosità” o altri artefatti. In alternativa si possono generare, con appositi pacchetti disponibili in Python, delle onde quadre o triangolari “modello” e verificare quantitativamente la differenza fra valori dell'onda modello e dell'onda creata per espansione di Fourier, per esempio usando la somma dei residui quadrati come indicatore della qualità di riproduzione.

III. SOTTO-ESERCIZIO 1: QUADRA E TRIANGOLARE

Il primo esercizio proposto è molto semplice: si tratta di ricostruire delle forme d'onda quadra e triangolare usando i coefficienti di Eqs. 12, 13. Può essere interessante osservare le forme d'onda corrispondenti a un numero via via crescente di iterazioni, cioè di elementi considerati nella somma di seni o coseni. La Fig. 3 mostra il risultato da me ottenuto per l'onda quadra con un numero crescente di iterazioni (il parametro n che compare in legenda). Notate che le forme d'onda sono graficate in funzione del tempo espresso in unità di periodo T : infatti in questo esercizio non è rilevante il valore fisico della frequenza. Inoltre, ovviamente, l'ampiezza ha unità arbitraria (il valore picco-picco è 1 [arb. un.]). Si vede come nel caso dell'onda quadra siano necessarie almeno alcune centinaia di iterazioni per ottenere la forma desiderata. Questo è dovuto al fatto che la forma d'onda quadra, con i suoi fronti di salita e discesa molto ripidi, non può essere riprodotta sommando su un numero troppo piccolo di componenti. Notate che il numero di componenti necessarie dipende, in parte, dal numero di punti che costituiscono la forma d'onda, ovvero l'array, che nel mio esempio è limitato a 1000 (non si notano evidenti problemi di aliasing nonostante il numero relativamente ridotto di punti).

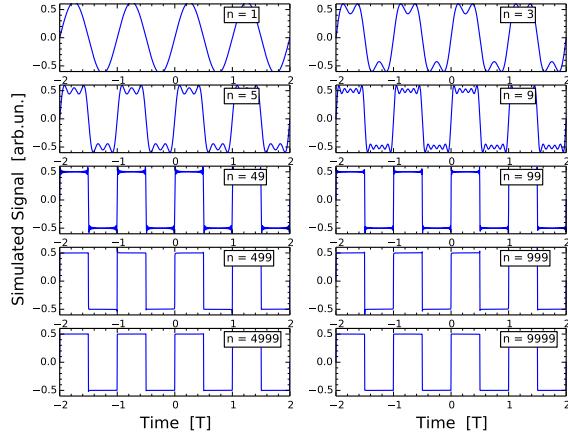


Figura 3. Sviluppo in serie di Fourier per un'onda quadra: i vari grafici si riferiscono a un numero n via via crescente di iterazioni, come indicato in legenda. Gli array graficati sono composti da 1000 punti.

La Fig. 4 si riferisce invece all'onda triangolare: qui, grazie all'assenza di ripidi fronti d'onda (in matematica derivate che tendono a divergere e non sono continue), è sufficiente un numero di iterazioni di poche decine per ottenere un risultato soddisfacente. In ogni caso l'efficienza del software permette di tenere alto il numero di iterazioni senza pregiudicare (troppo) la rapidità con cui i cicli vengono eseguiti.

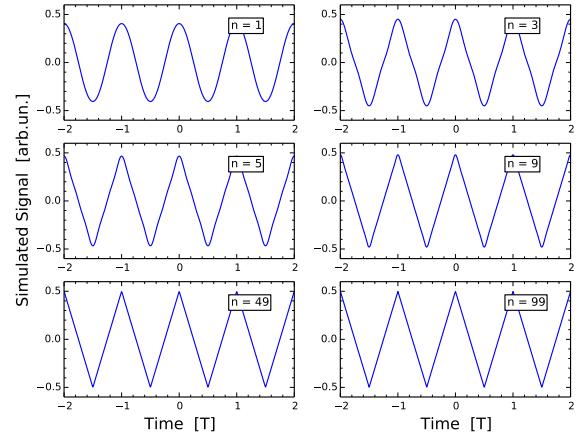


Figura 4. Sviluppo in serie di Fourier per un'onda triangolare: i vari grafici si riferiscono a un numero n via via crescente di iterazioni, come indicato in legenda. Gli array graficati sono composti da 1000 punti.

IV. SOTTO-ESERCIZIO 2: PINNA DI SQUALO

L'esercizio 2 qui proposto fa riferimento all'esperienza pratica in cui si osservava l'uscita di un integratore al cui ingresso era inviata un'onda quadra: all'aumentare della frequenza l'uscita tendeva ad assomigliare sempre più a un'onda triangolare, con un vasto regime intermedio in cui il segnale assumeva le sembianze di una sorta di pinna di squalo. Questa forma d'onda è stata acquisita con Arduino compatibilmente con il rate di campionamento disponibile, per varie frequenze f scelte arbitrariamente in un intervallo di oltre una decade.

Nell'esperimento era $C = 1 \mu\text{F}$ (con tolleranza $\pm 10\%$) e $R = (3.28 \pm 0.03)$ kohm (misurata con multmetro digitale), per cui la frequenza di taglio attesa era $f_T = (49 \pm 5)$ Hz. Arduino era regolato in modo da campionare a intervalli di durata nominale $\Delta t = 100 \mu\text{s}$ e i record erano acquisiti usando la combinazione di sketch e script `synclong2016`, in modo da ottenere 2048 coppie di punti (tempo e $V_{out}(t)$, quest'ultima lasciata in unità digitalizzate). Per poter impiegare Arduino il generatore di funzioni, impostato su onda quadra (di ampiezza $V_{in} \simeq 4.6$ Vpp), era stato regolato in modo da introdurre un offset che rendesse sempre positivo il segnale in ingresso e in uscita. Infatti in un circuito integratore la componente continua del segnale passa da ingresso a uscita.

La Fig. 5 mostra alcuni esempi dei record acquisiti a diverse frequenze, secondo quanto indicato in legenda. Per esigenze di chiarezza, i grafici sono stati disegnati usando diverse scale per gli assi: in particolare, l'asse orizzontale si estende per 4 periodi, cioè la scala va dallo zero dei tempi (ricordiamo che l'acquisizione è sincrona, per cui lo zero corrisponde sempre, entro l'incertezza, a un fronte di discesa dell'onda quadra in ingresso) al valore $4/f$, con f frequenza impostata e letta sul frequenzimetro del gene-

ratore di funzioni. Poiché l'intervallo di campionamento nominale è sempre lo stesso, la “densità” dei punti rappresentati, cioè il numero di dati effettivamente riportati nei grafici, dipende dalla frequenza, come è facile osservare. Inoltre, poiché anche l'offset del generatore è stato deliberatamente lasciato inalterato, i grafici a frequenze più alte, dove la componente alternata è attenuata, non partono dallo zero dell'asse verticale.

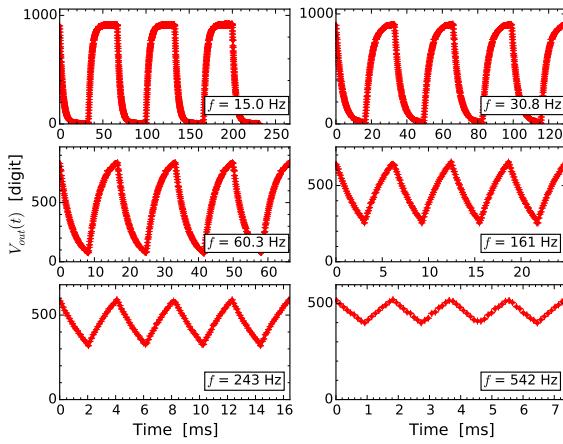


Figura 5. Dati acquisiti sperimentalmente come descritto nel testo: nei vari pannelli sono rappresentate (con punti e barre di errore “convenzionali”) i record $V_{out}(t)$ vs t registrati per diversi valori della frequenza, come indicato in legenda (per semplicità grafica, si omette l'indicazione precisa della frequenza e dell'incertezza sulla sua misura, che è più piccola dell'ultima cifra riportata). Si notino le diverse scale orizzontali e verticali.

Gli andamenti temporali osservati possono facilmente essere ricostruiti in maniera quantitativa usando il “metodo simbolico” e l’espansione di Fourier in seni e coseni.

Grazie all'applicazione del metodo simbolico, sappiamo che un integratore *alimentato con un'onda sinusoidale* di frequenza f dà luogo a:

1. un'guadagno, o attenuazione, $A(f) = 1/\sqrt{1 + (f/f_T)^2}$;
2. uno sfasamento $\Delta\phi = \arctan(-f/f_T)$.

Attenuazione e sfasamento agiscono indipendentemente su tutte le componenti dell'onda quadra supposta in ingresso al circuito, avendo per ogni componente $f = f_k = \omega_k/(2\pi)$. In altre parole, si può agevolmente “simulare” la forma d'onda $w(t)$ in uscita dal circuito sommando le componenti armoniche dell'onda quadra, cioè costruita con i coefficienti di Eq. 12, moltiplicando l'ampiezza di ogni componente per la dovuta attenuazione e mettendo nell'argomento delle armoniche di Fourier il dovuto sfasamento.

In altre parole avremo:

$$A_k = \frac{1}{\sqrt{1 + (\omega_k/\omega_T)^2}} \quad (14)$$

$$\Delta\phi_k = \arctan(-\omega_k/\omega_T) \quad (15)$$

$$w(t) = \sum_{k=1}^n c_k A_k \sin(\omega_k t + \Delta\phi_k), \quad (16)$$

con k dispari e c_k dato da Eq. 12. Ci occuperemo in seguito della presenza di un offset non nullo, cioè del termine $a_0 \neq 0$ in Eq. 12, e dei problemi di fase iniziale, che nell'acquisizione sperimentale o è random (se si usa la combinazione sketch/script `ardu2016`) oppure è determinata dalla strategia di sincronizzazione del generatore di funzioni con Arduino (se si usa, come in questo esempio, la combinazione `synclong2016`).

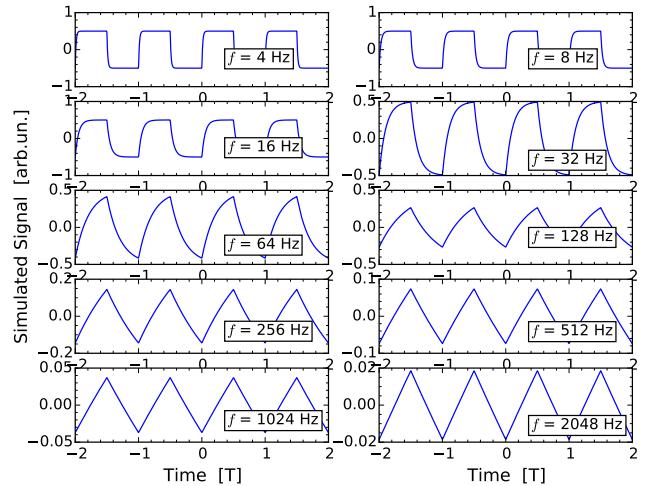


Figura 6. “Simulazione” dell'integratore descritto nel testo ($f_T = 48.6$ Hz) con all'ingresso un'onda quadra. I vari pannelli si riferiscono a diverse frequenze, come in legenda. Fate attenzione alla scala orizzontale, in unità di periodi $T = 1/f$, e alla scala verticale, che cambia di grafico in grafico. Gli array graficati sono composti da 1000 punti e realizzati sommando su 1000 iterazioni.

La Fig. 6 mostra l'esito della “simulazione” supponendo $f_T = 48.6$ Hz, cioè il valore nominale (senza incertezza) dell'integratore realizzato sperimentalmente. I vari pannelli mostrano il segnale simulato per diverse frequenze, come indicato in legenda: le frequenze scalano di un fattore due da un grafico al successivo e non corrispondono ai valori sperimentali, poiché questa figura ha lo scopo principale di illustrare l'evoluzione della forma d'onda da quadrata a triangolare all'aumentare della frequenza. Fate attenzione al fatto che le scale orizzontali sono in unità di periodo $T = 1/f$ e quelle verticali sono diverse per i vari grafici.

Il risultato è in accordo con le osservazioni: a basse frequenze, $f < f_T$, la forma d'onda non viene praticamente

modificata e all'uscita si ritrova il segnale inviato all'ingresso. Aumentando la frequenza si nota una deformazione della forma d'onda, accompagnata da un'attenuazione dell'ampiezza e da uno sfasamento, che determinano, per $f >> f_T$, un'uscita di forma pressoché triangolare, evidentemente sfasata di circa $|\pi/2|$ rispetto all'ingresso. Dal punto di vista qualitativo, già per $f \sim 10f_T$ la forma d'onda in uscita è ben approssimata da un andamento triangolare.

La Fig. 7 mostra, sovrapposti, i risultati sperimentali per tre diverse frequenze e quelli ottenuti da simulazioni. Come già sottolineato ci sono alcune differenze tra l'uscita della simulazione, realizzata come descritto sopra, e i dati acquisiti da Arduino:

1. c'è un fattore di fase costante, che vale circa $\Delta\Phi \simeq \pi$, causato dalla strategia di sincronizzazione: i dati sono infatti acquisiti a partire da un istante nel quale l'onda quadra è nel suo fronte di discesa;
2. c'è un offset, cioè $a_0 \neq 0$ in Eq. 12, dovuto alla circostanza che Arduino accetta in ingresso solo d.d.p. positive (o nulle) e che pertanto un offset era stato introdotto in ingresso;
3. c'è un fattore di scala, cioè l'onda quadra in ingresso, e di conseguenza quella in uscita, non hanno ampiezza picco-picco unitaria.

Le tre differenze possono essere facilmente sanate implementando *a mano* (e “a occhio”) alcuni accorgimenti, ad esempio:

1. la forma d'onda simulata viene graficata dopo averla traslata di un semiperiodo (che corrisponde a $\Delta\Phi = \pi$), ovvero del valore necessario nel caso di acquisizioni asincrone;
2. alla forma d'onda simulata viene aggiunto un termine costante desunto dal valore medio (su un periodo) delle acquisizioni;
3. la forma d'onda simulata viene moltiplicata per un fattore che tiene conto dell'ampiezza del segnale acquisito, desunto dalle registrazioni a bassa frequenza (dove l'attenuazione è trascurabile).

Infine, è ovvio che, allo scopo di migliorare qualitativamente l'accordo tra simulazione e dati sperimentali, il valore della frequenza di taglio f_T impiegata nella simulazione può essere aggiustato all'interno dell'incertezza con cui esso è noto: le simulazioni di figura sono state calcolate ponendo $f_T = 46$ Hz, compatibile con il valore nominale dato dalla misura di R e dalla conoscenza di C . Naturalmente, la procedura a mano, o “a occhio”, potrebbe essere sostituita da un (complicato) best-fit finalizzato a individuare i valori dei parametri attraverso minimizzazione del χ^2 : tuttavia esso non è certamente necessario. Infatti, come si può facilmente osservare, la simulazione ottenuta a mano e “a occhio” riproduce in

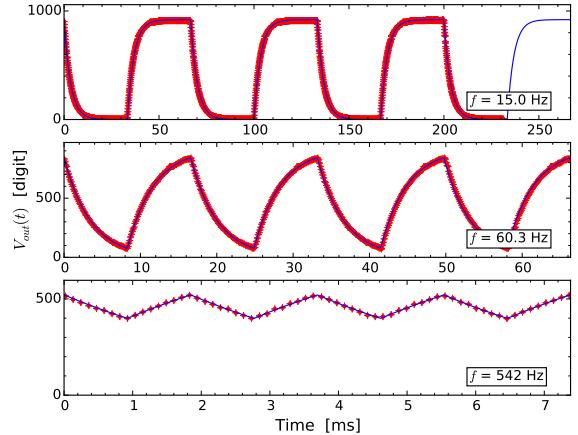


Figura 7. Dati acquisiti sperimentalmente (punti rossi con barre di errore) e simulazioni (linee continue blu) per tre diverse frequenze. Dettagli sull'acquisizione e sulla realizzazione delle simulazioni sono dati nel testo. Si noti, nel pannello superiore, la diversa durata del record sperimentale e di quello simulato.

modo qualitativamente soddisfacente i dati sperimentali e ciò è sufficiente per i nostri scopi.

Se il tempo le permette e la vostra curiosità vi spinge a farlo, ci sono numerose altre osservazioni sperimentali nel dominio del tempo, cioè forme d'onda visualizzate all'oscilloscopio, che potete divertirvi a ricostruire. Per esempio, alcuni di voi hanno mostrato un certo stupore nel verificare come la forma d'onda in ingresso (la V_{in} , per usare la terminologia dell'esercitazione pratica) venisse “distorta” dalla presenza del circuito integratore. Questa “distorsione” nel nostro modello è dovuta al carattere reale del generatore di forme d'onda, cioè alla presenza di una resistenza interna ($r_G = 50$ ohm, nominale) non nulla e alla conseguente caduta di potenziale. Essa dipende dall'intensità di corrente, I_ω nell'approccio fasoriale, che circola nella maglia costituita dalla serie RC. Quindi *dovrebbe* essere possibile ricostruire numericamente questa caduta di potenziale, e di conseguenza il segnale osservato all'oscilloscopio, in funzione dei parametri di operazione. Inoltre potreste estendere questo esercizio considerando in ingresso una forma d'onda triangolare, oppure esaminare un derivatore invece di un integratore.

V. SOTTO-ESERCIZIO 3: AMPIEZZA DELLA FORMA D'ONDA INTEGRATA

Questo sotto-esercizio è finalizzato a determinare l'andamento dell'ampiezza in uscita dall'integratore in funzione della frequenza f , sempre supponendo una forma d'onda quadra in ingresso. Nell'illustrazione dell'integratore studiato nel dominio del tempo abbiamo accennato alla circostanza che il segnale in uscita, in condizioni di integratore ben funzionante, può essere approssimato

con il primo ordine dello sviluppo di Taylor dell'andamento esponenziale contenuto nella soluzione analitica della carica/scarica del condensatore. Tuttavia questa descrizione è necessariamente approssimativa (appunto, al primo ordine!). D'altra parte il modello dell'integratore, ovvero del filtro passa-basso, che abbiamo costruito per l'analisi nel dominio delle frequenze permette di predire facilmente tale andamento *solamente* nel caso di *onde sinusoidali* attraverso la relazione, già ampiamente citata, $A(f) = 1/\sqrt{1 + (f/f_T)^2}$. Lo sviluppo in armoniche di Fourier può modificare questo andamento a causa della diversa attenuazione a cui sono sottoposte le varie armoniche.

Nell'esperimento svolto da me il guadagno $A(f)$ è stato misurato per l'integratore costruito come specificato in precedenza. A questo scopo è stato calcolato il rapporto tra le ampiezze picco-picco (più facili da individuare) del segnale in uscita e di quello in ingresso, entrambi monitorati con i due canali dell'oscilloscopio. Grazie all'elevata resistenza in ingresso di questo strumento è stato supposto, senza dimostrazione (questo argomento sarà trattato un po' più in dettaglio altrove), che l'oscilloscopio perturbasse in maniera trascurabile il circuito. Analogamente è stato riservato alla resistenza di uscita del generatore, che è stata supposta trascurabile.

La simulazione è stata fatta girare per diversi valori della frequenza f dentro un intervallo di alcune decadi, grosso modo corrispondente a quello esplorato sperimentalmente. Nelle simulazioni si è supposta un'ampiezza picco-picco unitaria in ingresso, mentre quella del segnale in uscita è stata dedotta numericamente come differenza tra valore massimo e valore minimo del segnale simulato. Per ottenere previsioni affidabili si è fatto in modo che le forme d'onda simulate non presentassero "spigolosità" artificiose dovute a sotto-campionamento. Inoltre si è scelta $f_T = 46$ Hz sulla base dei risultati precedenti,

La Fig. 8 mostra i dati sperimentali (punti e barre di errore rossi) con sovrapposti i risultati della simulazione (linea continua blu) e l'andamento previsto per la forma d'onda sinusoidale (linea tratteggiata grigia). È evidente come il comportamento sperimentale sia diverso dalle previsioni per il caso sinusoidale e anche come esso sia qualitativamente ben riprodotto dalla simulazione.

VI. SOTTO-ESERCIZIO 4: INTEGRATORE + DERIVATORE

Questo sotto-esercizio richiede di considerare il circuito composto da integratore e derivatore in cascata realizzato nell'esperienza pratica. Lo scopo è anche in questo caso quello di "simulare" le forme d'onda dei segnali in uscita da integratore e derivatore, rispettivamente V_A e V_B secondo la nomenclatura dell'esercitazione pratica, e di verificare il rapporto tra le loro ampiezze e l'ampiezza dell'onda quadra in ingresso, V_{in} . Quando integratore e derivatore funzionano come devono, l'uscita A avrà forma triangolare e l'uscita B quadra.

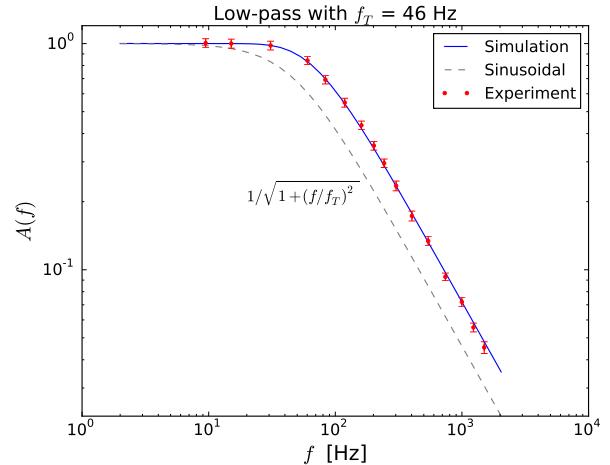


Figura 8. Guadagno, o attenuazione, A in funzione della frequenza f determinato sperimentalmente come descritto nel testo (punti e barre di errore rossi), ottenuto dalla "simulazione" descritta nel testo (la linea continua blu raccorda diversi punti simulati, equispaziati logaritmicamente nell'intervallo di interesse) e dalla previsione per onde sinusoidali (linea tratteggiata grigia), calcolata secondo la funzione riportata nel grafico. Per le previsioni si è supposto $f_T = 46$ Hz, compatibile con il valore sperimentale.

Per semplicità, in questo esercizio *non si considerano gli effetti dovuti al collegamento in cascata dei due sotto-circuiti*, né quelli legati alla presenza di resistenze interne; gli effetti dovuti al matching tra diversi sotto-circuiti e alle resistenze interne potranno essere considerati in altra sede. In buona sostanza, si suppone qui di avere un integratore che ha frequenza di taglio f_{TA} con in cascata un derivatore che ha frequenza di taglio f_{TB} .

Nel mio esempio ho supposto $f_{TA} = 50$ Hz e $f_{TB} = 25$ kHz; naturalmente voi siete invitati a usare i valori effettivi dei circuiti che avete montato e testato. Per determinare l'uscita in A (uscita dell'integratore) $w_A(t)$ ho impiegato in pratica le stesse relazioni di Eq. 14

$$A_{A,k} = \frac{1}{\sqrt{1 + (\omega_k/\omega_{TA})^2}} \quad (17)$$

$$\Delta\phi_{A,k} = \arctan(-\omega_k/\omega_{TA}) \quad (18)$$

$$w_A(t) = \sum_{k=1}^n c_k A_{A,k} \sin(\omega_k t + \Delta\phi_{A,k}), \quad (19)$$

dove k è dispari e c_k è dato dall'Eq. 12.

Come ben sapete, l'ulteriore stadio di derivazione, con frequenza di taglio f_{TB} , introduce sulle diverse armoniche:

1. un ulteriore *guadagno, o attenuazione*, $A(f) = 1/\sqrt{1 + (f_{TB}/f)^2}$;
2. un ulteriore *sfasamento* $\Delta\phi = \arctan(f_{TB}/f)$.

Queste ulteriori modifiche delle armoniche agiscono "in cascata": l'attenuazione andrà a moltiplicare l'ampiezza

delle componenti armoniche già attenuate dall'integratore, e lo sfasamento andrà a sommarsi allo sfasamento prodotto dall'integratore. Detta $w_B(t)$ la forma d'onda in uscita da B, si può quindi scrivere

$$A_{B,k} = \frac{1}{\sqrt{1 + (\omega_{TB}/\omega_k)^2}} \quad (20)$$

$$\Delta\phi_{B,k} = \arctan(\omega_{TB}/\omega_k) \quad (21)$$

$$w_B(t) = \sum_{k=1}^n c_k A_{A,k} A_{B,k} \sin(\omega_k t + \Delta\phi_{A,k} + \Delta\phi_{B,k}) \quad (22)$$

dove $A_{A,k}$ e $\Delta\phi_{A,k}$ sono quelli determinati in Eqs. 17, 18. Notate che, come è ovvio, per le frequenze a cui integratore e derivatore si comportano come si deve i due sfasamenti si annullano a vicenda, tendendo rispettivamente a $-\pi/2$ (integratore) e $\pi/2$ (derivatore), per cui in uscita da B si ritrova un'onda quadra in fase con quella in ingresso e ovviamente attenuata.

Le Figs. 9, 10 mostrano i risultati, cioè le forme d'onda "simulate" in uscita rispettivamente da A e da B. Il range di frequenze f considerato parte da f_{TA} e arriva a f_{TB} . Notate che anche in queste figure l'asse orizzontale è espresso in unità di periodo T , mentre l'asse verticale cambia di grafico in grafico.

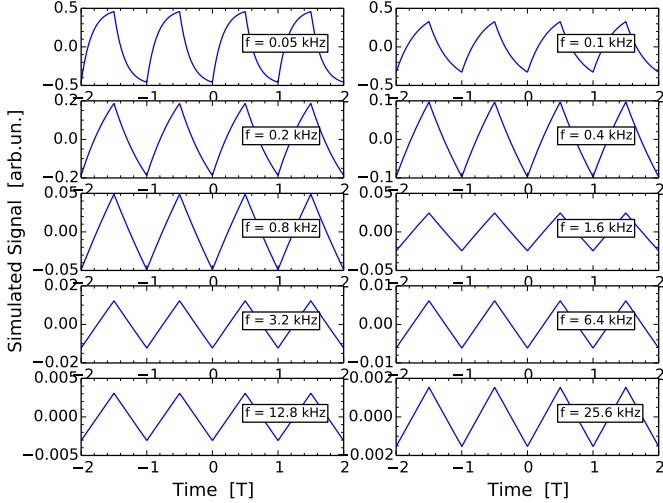


Figura 9. "Simulazione" del segnale in uscita dall'integratore descritto nel testo ($f_{TA} = 50$ Hz) con all'ingresso un'onda quadra. I vari pannelli si riferiscono a diverse frequenze, come in legenda. Fate attenzione alla scala orizzontale, in unità di periodi $T = 1/f$, e alla scala verticale, che cambia di grafico in grafico. Gli array graficati sono composti da 1000 punti e realizzati sommando su 1000 iterazioni.

Vediamo cosa succede per la forma d'onda simulata in uscita dal derivatore (uscita B, Fig 10). Questo sottocircuito funziona come derivatore solo per $f \ll f_{TB}$. In effetti a basse frequenze il derivatore fa il suo mestiere, ma agisce su una forma d'onda che è praticamente quadra, dato che a queste frequenze l'integratore non integra

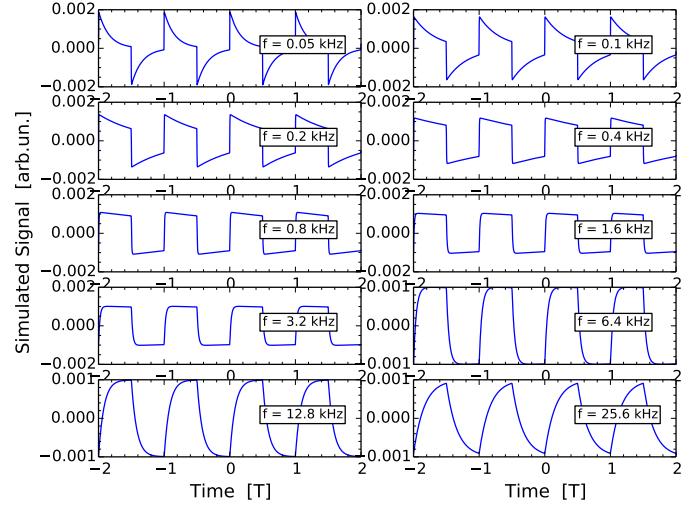


Figura 10. "Simulazione" del segnale in uscita dalla cascata integratore+derivatore descritta nel testo ($f_{TB} = 25$ kHz) con all'ingresso un'onda quadra. I vari pannelli si riferiscono a diverse frequenze, come in legenda. Fate attenzione alla scala orizzontale, in unità di periodi $T = 1/f$, e alla scala verticale, che cambia di grafico in grafico. Gli array graficati sono composti da 1000 punti e realizzati sommando su 1000 iterazioni.

"abbastanza": nel segnale ricostruito numericamente si vedono chiaramente le "tracce" della derivata di un'onda quadra, che per la matematica è costituita da una successione di funzioni tipo delta, separate temporalmente da mezzo periodo e dirette alternativamente verso l'alto e verso il basso. C'è poi un intervallo di frequenze (per esempio per i grafici corrispondenti a $f = 1.6$ kHz e $f = 3.2$ kHz) in cui l'uscita B riporta una forma d'onda quadra: qui sia l'integratore che il derivatore fanno il loro dovere. Se la frequenza viene ulteriormente aumentata si esce dalla condizione $f \ll f_{TB}$ e la forma d'onda in uscita dal derivatore tende a essere inalterata rispetto a quella che si trova al suo ingresso: dunque essa tende ad essere triangolare, o, se preferite, a pinna di squalo.

Soffermiamoci anche qui a esaminare i guadagni, o attenuazioni, tra ampiezza picco-picco in uscita B, cioè del segnale $V_B(t)$, e ampiezza picco-picco in ingresso *all'intero circuito*, cioè $V_{in}(t)$. Ricordiamo che, in questo caso, il valore atteso per *onde sinusoidali* è $A_{B,att} = f_{TA}/f_{TB}$ *indipendente dalla frequenza*. Per la scelta delle frequenze di taglio operata nel nostro esempio si ha $A_{B,att} = 2 \times 10^{-3}$: dato che il segnale in ingresso è assunto avere un'ampiezza picco-picco unitaria (in unità arbitrarie), questo vuol dire che il segnale in uscita da B deve avere un'ampiezza di 2×10^{-3} [arb. un.]. Si vede subito da Fig. 10 che questo è approssimativamente quanto si ottiene con la simulazione, in particolare nel range in cui integratore e derivatore funzionano come tali. Anche qui l'interpretazione è piuttosto semplice: presi separatamente, i guadagni, o attenuazioni, di integratore e derivatore di-

pendono dalla frequenza di lavoro in un modo che non segue la semplice aspettativa riferita a onde sinusoidali. Tuttavia i due circuiti sotto-attenuano e sovra-attenuano le varie componenti armoniche in un modo che finisce per compensarsi. Alla fine, l'attenuazione complessiva risulta indipendente dalla frequenza di lavoro e in ragionevole accordo con le attese. Questo accordo peggiora alle frequenze in cui integratore e derivatore non si comportano pienamente come tali.

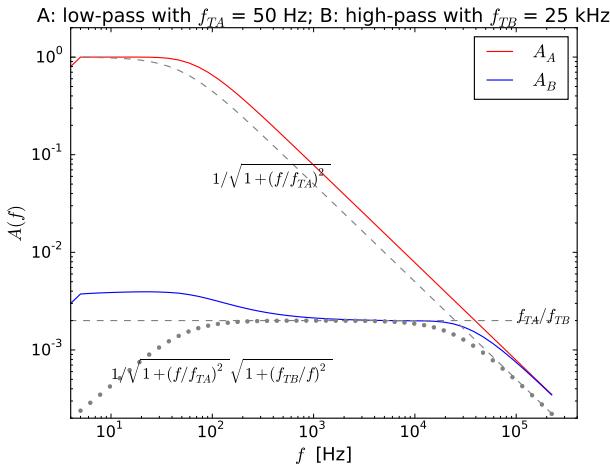


Figura 11. “Simulazione” dei guadagni, o attenuazioni, A_A e A_B in uscita dall'integratore e dalla cascata integratore+derivatore descritta nel testo in funzione della frequenza di lavoro e per i valori di frequenza di taglio riportati nel testo e nel titolo. Le curve tratteggiate e punteggiate rappresentano gli andamenti attesi per onde sinusoidali (le funzioni corrispondenti sono scritte sul grafico). Il grafico si riferisce a 50 valori di frequenza distinti e equispaziati logaritmicamente nell'intervallo considerato (la linea continua raccorda i diversi punti).

Il comportamento complessivo in termini di ampiezza del segnale in uscita è riassunto in Fig. 11: essa è stata costruita come la Fig. 8, cioè eseguendo un ciclo in un range di frequenze (equispaziate logaritmicamente) e deducendo l'ampiezza picco-picco del segnale in uscita (dal sotto-circuito A, linea continua rossa, o dalla cascata dei sotto-circuiti A e B, linea continua blu) come differenza tra valore massimo e minimo della forma d'onda simulata. Si noti che in questo caso la definizione impiegata potrebbe non essere del tutto affidabile, a causa delle spigolosità di alcune forme d'onda simulate (si vedano i primi pannelli di Fig. 10). Il risultato mostra che la cascata integratore+derivatore ha un intervallo di frequenza nel quale l'attenuazione complessiva per l'onda quadra in ingresso è simile a quanto previsto nel caso sinusoidale (il rapporto f_{TA}/f_{TB} indipendente dalla frequenza, indicato con una linea tratteggiata grigia nel grafico). Al di fuori di questo range l'attenuazione segue un andamento che tende a discostarsi da quello previsto per onde sinusoidali

(il prodotto $A_A A_B$ rappresentato dalla linea punteggiata grigia nel grafico): in particolare, e anche a causa della definizione di ampiezza picco-picco impiegata, a frequenze basse il guadagno è maggiore rispetto a quanto atteso per le onde sinusoidali. Per il resto, è ovvio che il comportamento dell'integratore, che già abbiamo analizzato in Fig. 8 (in questa figura esso è rappresentato da una linea tratteggiata grigia), gioca un ruolo nel determinare il guadagno complessivo della cascata.

Nell'esercitazione pratica almeno qualcuno o qualcuna tra voi ha acquisito i dati rilevanti per l'ampiezza (io non l'ho fatto!). Dunque dovrebbe essere possibile ricostruire numericamente gli andamenti sperimentali. Infine, anche in questo caso potreste estendere l'analisi simulata ad altre situazioni, per esempio verificare cosa succede se in ingresso supponete di avere un'onda triangolare, oppure se impiegate altri valori per resistenze e condensatori, o magari se scambiate tra di loro integratore e derivatore (qui l'aspettativa è banale, ma comunque interessante).

VII. SOTTO-ESERCIZIO 5: ACCOPPIAMENTO AC IN INGRESSO ALL'OSCILLOSCOPIO

Come ben sapete, l'accoppiamento AC in ingresso all'oscilloscopio serve per cancellare la componente continua del segnale visualizzato. Questa possibilità è estremamente utile nei (frequenti) casi in cui il segnale è costituito da una (piccola) componente variabile nel tempo, che contiene l'informazione di interesse, sovrapposta a un piedistallo costante.

Questa modalità di operazione dell'oscilloscopio può essere modellata supponendo la presenza di un condensatore di capacità C_{AC} in serie al segnale: una volta carico, esso impedisce il passaggio di corrente, sopprimendo la componente continua del segnale. Assieme alla resistenza di ingresso dell'oscilloscopio (tipicamente per noi è $r_{osc} = 1$ Mohm nominale [1]), montata in parallelo al segnale, cioè tra segnale e linea di massa, o terra, il condensatore realizza di fatto un circuito derivatore RC, ovvero un filtro passa-alto.

Una conseguenza eclatante dell'accoppiamento AC è la forte “distorsione” di una forma d'onda quadra a bassa frequenza, che prende una forma caratteristica in cui i tratti orizzontali assumono andamenti di tipo esponenziale. Questa distorsione può essere facilmente ricostruita numericamente applicando attenuazione e sfasamento a un'onda quadra scritta in componenti di Fourier. La Fig. 12 simula la visualizzazione all'oscilloscopio di un'onda quadra di frequenza $f = 40$ Hz sottoposta all'azione di un filtro passa-alto con $f_T = 10$ Hz [2].

Purtroppo in questo caso non riesco a confrontare simulazione e osservazione sperimentale, ma magari a voi sarà possibile farlo, per esempio prendendo con il telefonino uno screenshot dello schermo dell'oscilloscopio (cosa che io non ho fatto).

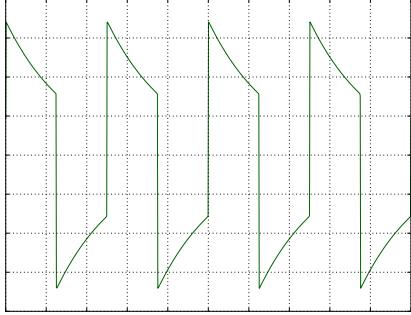


Figura 12. “Simulazione” della visualizzazione di un’onda quadra di frequenza $f = 40$ Hz e ampiezza 0.5 V con un oscilloscopio accoppiato in AC. L’oscilloscopio è regolato in modo da avere una velocità di sweep di 10 ms/div e un’amplificazione di 200 mV/div: osservate come l’ampiezza picco-picco visualizzata in AC sia maggiore di quella effettiva della forma d’onda a causa dell’operazione di derivata temporale. Il derivatore in ingresso all’oscilloscopio è supposto avere $f_T = 10$ Hz. È chiaro che la forma d’onda simulata potrebbe differire da quanto osservato all’oscilloscopio: è infatti possibile che sullo schermo dell’oscilloscopio i segmenti verticali, che rappresentano brusche variazioni di d.d.p. in funzione del tempo, non siano (ben) visibili a causa del troppo scarso deposito di energia sullo schermo dell’oscilloscopio, conseguente al rapido movimento del pennello elettronico.

- [1] Come sapete, un modello migliore per lo stadio di ingresso dell’oscilloscopio prevede il parallelo tra r_{osc} e un piccolo condensatore $C_{osc} = 25$ pF (nominali). La presenza del condensatore è necessaria anche per modellare capacità parassite inevitabilmente presenti nei connettori e nei circuiti dell’oscilloscopio (due conduttori posti uno vicino all’altro e interessati da correnti formano praticamente sempre un condensatore!). Dunque lo stadio di ingresso ha una *impedenza di ingresso* data da questo parallelo. I moduli delle impedenze dei due componenti in parallelo

diventano paragonabili per $\omega \sim 4 \times 10^4$ rad/s. Quindi trascurare la presenza del condensatore C_{osc} , come fatto nel testo, è a rigore giustificato tanto più quanto più la frequenza angolare di lavoro è inferiore rispetto a questo valore.

- [2] Supponendo che la frequenza di taglio del derivatore realizzato da C_{AC} sia determinata da r_{osc} , ovvero $f_T = 1/(2\pi r_{osc} C_{osc})$, si ottiene $C_{osc} \sim 4$ nF.

Giunzione p-n e diodo al silicio

francesco.fuso@unipi.it

(Dated: version 4 - FF, 2 gennaio 2020)

Il funzionamento delle *giunzioni bipolari* fatte di materiali semiconduttori drogati, per esempio silicio, è un argomento tradizionale di gigantesco interesse per le applicazioni in elettronica e optoelettronica. La breve trattazione che viene qui presentata, in cui si evita accuratamente di discutere o citare aspetti al di fuori delle conoscenze di fisica classica (elettrostatica), è finalizzata a interpretare alcuni rilevanti aspetti dell'operazione dei diodi a giunzione bipolare.

I. INTRODUZIONE

In termini generali, si parla di *giunzione* quando si ha un'interfaccia, cioè una superficie di separazione, tra due materiali con caratteristiche elettroniche, cioè di *trasporto elettrico*, diverse. Nella tecnologia attuale hanno un ruolo importantissimo tante differenti tipologie di giunzione, che coinvolgono materiali con caratteristiche disparate (isolanti, conduttori, semiconduttori, superconduttori, inorganici, organici, nanostrutture, etc.). Qui intendiamo brevemente descrivere il comportamento della giunzione realizzata fra due materiali *semiconduttori*; in particolare, tra i tanti semiconduttori di uso comune nella tecnologia attuale, spesso composti da leghe binarie e ternarie di elementi appartenenti alle colonne tra la II e la VI della tavola periodica, qui ci limitiamo a considerare pezzi di silicio *drogati* in modo p e n. Questo tipo di giunzione, oltre ad avere una fondamentale rilevanza storica (la sua realizzazione su larga scala sancì la nascita dell'elettronica a stato solido, ormai oltre 50 anni fa), è di grandissima importanza e diffusione nei dispositivi per l'elettronica e l'opto-elettronica e, per esempio, è quella integrata negli ordinari diodi a stato solido usati in laboratorio.

Come sarà chiaro nel seguito, questo tipo di giunzione è di fatto realizzata tra due regioni dello stesso materiale nelle quali sono presenti in maggioranza cariche libere (*portatori di carica*, secondo la nomenclatura di uso comune in questo ambito) di polarità positiva e negativa, rispettivamente nei pezzi di silicio drogati p e n. Questo dà ovviamente ragione alla differenza di proprietà di trasporto elettrico delle due parti. Per cominciare, però, occorre discutere le proprietà elettriche (o elettroniche) dei materiali semiconduttori "puri", cioè non drogati: tali proprietà si definiscono *intrinseche*. La discussione sarà svolta in maniera molto qualitativa e limitando al massimo ogni accenno agli aspetti di meccanica quantistica che vi sono coinvolti.

II. SEMICONDUTTORI

Dal punto di vista delle proprietà di trasporto elettrico un semiconduttore (intrinseco, cioè non drogato) è banalmente un materiale conduttore con resistività piuttosto alta: nel caso del silicio si ha una resistività

$\rho_C \simeq 2.5 \times 10^3$ ohm m a temperatura ambiente, da confrontare con $\rho_C \sim 10^{-8}$ ohm m tipica dei metalli (rame, argento, oro, etc.). Questa proprietà è valsa proprio la denominazione di semiconduttori a tali materiali. È noto da tempo che essi mostrano anche un'altra proprietà distintiva: all'aumentare della temperatura la loro resistività tende a diminuire (comportamento *NTC*, cioè con coefficiente di temperatura negativo), a differenza di quanto succede normalmente nei conduttori metallici, come avevamo illustrato facendo riferimento al "modello di Drude".

Per il silicio, e per altri semiconduttori elementari del gruppo IV della tavola periodica, queste proprietà sono conseguenza piuttosto diretta della loro struttura microscopica. Tali materiali, infatti, si presentano facilmente con una struttura cristallina in cui ogni atomo è coordinato ("legato") con quattro atomi adiacenti. Idealmente, una configurazione di questo tipo renderebbe il materiale un perfetto isolante, dato che nessuna carica sarebbe libera di muoversi, cioè potrebbe agire da *portatore di carica*. Infatti i quattro elettroni "di valenza" di un atomo sono compartecipati con quattro elettroni di quattro atomi adiacenti a formare dei "legami" che in chimica si definirebbero "covalenti", rimanendo localizzati attorno al nucleo atomico (o ione) che li ha messi a disposizione.

A. Generazione

Nella realtà ci sono diversi motivi per cui, invece, la densità dei portatori di carica in un semiconduttore cristallino non è nulla. A parte quelli legati alla presenza di imperfezioni nel cristallo o di impurezze, il fatto stesso che il sistema considerato si trovi a una temperatura diversa da zero implica un probabilità non nulla che alcuni dei legami covalenti fra atomi contigui non si realizzino. La differenza tra semiconduttori e dielettrici propriamente detti è proprio nella probabilità che alcuni legami non si formino, o vengano "annullati", già a temperatura ambiente. Questa probabilità è comparativamente maggiore per i semiconduttori, nei quali si ha quindi una densità non nulla di portatori di carica originati proprio dalla mancanza di questi legami.

Anche se, come promesso, non intendiamo assolutamente svolgere qui una trattazione quantistica, vale la pena di accennare brevemente alla circostanza che la de-

scrizione quantistica di un semiconduttore cristallino, come probabilmente studierete in futuro, prevede che questa sorta di rottura di legame sia ben descritta come “promozione”, o “transizione”, di un elettrone dallo stato localizzato (di legame, o valenza) a quello delocalizzato (di portatore di carica, o conduzione). Questa promozione avviene quando al sistema viene fornita energia sufficiente, al di sopra di una soglia che è relativamente bassa. Essa, che prende il nome di *energia di gap*, vale circa 1.1 eV nel caso del silicio a temperatura ambiente (si ricorda che $1 \text{ eV} \simeq 1.6 \times 10^{-19} \text{ J}$). Un sistema che si trova a temperatura ambiente ($T \sim 300 \text{ K}$) interagisce con un bagno termico con energia caratteristica $k_B T \simeq 1/40 \text{ eV}$, con k_B costante di Boltzmann, che segue una distribuzione di tipo Maxwell-Boltzmann. Pertanto la probabilità che sia disponibile energia sufficiente a superare il gap è non nulla, anche se molto bassa (l’andamento è di tipo esponenziale, con argomento proporzionale al rapporto tra l’energia di gap e quella termica caratteristica, o media).

In ogni caso la densità dei portatori di carica liberi in un semiconduttore intrinseco, che qui indichiamo con n , è bassa, da cui la conducibilità comparativamente minore che per i conduttori metallici. Infatti la densità di corrente in un conduttore, $\vec{j} = nq\vec{v}_d$, con q carica elementare e \vec{v}_d velocità classica di deriva, è direttamente proporzionale alla densità dei portatori di carica. Per tenere conto della circostanza che in nessun caso i portatori sono completamente liberi di muoversi (così come in un metallo, secondo il modello di Drude), in questo ambito si introduce una nuova grandezza μ , detta *mobilità*, definita da $\mu = |\vec{v}_d|/|\vec{E}|$, con \vec{E} campo elettrico applicato. Valori tipici della mobilità per il silicio intrinseco a temperatura ambiente sono dell’ordine di $10^2 - 10^3 \text{ cm}^2/(\text{V s})$ (notate l’unità di misura). Tali valori, combinati con la densità di portatori di carica tipica a temperatura ambiente, danno luogo alla resistività relativamente elevata propria dei semiconduttori. Inoltre è ovvio che, essendo il meccanismo di generazione di portatori di carica attivato termicamente, all’aumentare della temperatura si assiste a un aumento della densità di corrente e dunque, a parità di altre condizioni, a una diminuzione della resistività.

B. Elettroni (n) e lacune (p)

Evidentemente l’effetto di rendere disponibile al trasporto un elettrone in seguito alla “rottura del legame”, ovvero per promozione allo stato di conduzione è inerentemente accompagnato dall’assenza di un elettrone localizzato in un punto del cristallo. Quindi il meccanismo di cui stiamo trattando conduce alla *generazione di una coppia* di “entità”, rappresentate rispettivamente dall’elettrone e dalla sua assenza;, a cui daremo il nome di *lacuna*. È chiaro che questa assenza localizzata di un elettrone può essere colmata da un elettrone che proviene da un altro “legame rotto”, situato da qualche altra parte nel cristallo, che dunque si rende disponibile per ri-formare un legame atomico. Spesso a questo processo

si dà il nome di *ricombinazione*; secondo la meccanica quantistica, ad esso può essere associato un rilascio di energia che può essere dell’ordine dell’energia di gap.

Non ci stupisce che gli elettroni che si formano in seguito alla rottura del legame siano liberi di muoversi: in quanto particelle piccole e dotate di piccola massa, gli elettroni potranno muoversi nel cristallo, almeno finché non ricombinano. Dunque è piuttosto immediato identificare gli elettroni come portatori di carica negativa. Ora, immaginiamo che in un certo sito del nostro cristallo si generi una coppia elettrone lacuna, e che l’elettrone si muova in una certa direzione, per esempio verso destra, sotto l’effetto di un campo elettrico esterno applicato al sistema. Come già sottolineato, l’elettrone lascia una lacuna, cioè un’assenza localizzata nel sito di origine. Immaginiamo poi che, nel suo moto verso destra, l’elettrone incontri un sito, diverso da quello di partenza, in cui si trova un’altra assenza di elettrone, e qui si ricombini: in questo processo, della carica negativa si è spostata da sinistra a destra, dal sito dell’originaria generazione a quello della ricombinazione. Però possiamo facilmente renderci conto che è vera anche l’affermazione che della carica positiva si è spostata da destra verso sinistra, cioè dal sito della ricombinazione a quello dell’originaria generazione.

Anche se a muoversi possono essere soltanto gli elettroni, dal punto di vista elettrico il modello che stiamo usando prevede che nel semiconduttore ci siano *portatori di carica, cioè cariche libere, di ambo i segni*: il portatore di carica negativa continua a chiamarsi *elettrone*, a quello di carica positiva si dà il nome di *lacuna* (talvolta buca, o valenza). Assumendo come valido questo modello, la cui potenza ci sarà chiara fra breve, potremo dimenticarci dei meccanismi microscopici (classici) che abbiamo descritto finora e, appunto, concentrarci sul solo fatto che in un semiconduttore esistono portatori di carica negativi e positivi.

Vale la pena sottolineare che quanto stiamo qui affermando si riferisce a un meccanismo specifico per i semiconduttori. Infatti anche in un conduttore metallico è possibile associare al movimento di elettroni in una direzione e verso con il movimento nella stessa direzione e in verso opposto di cariche unitarie positive: la matematica consente di cambiare segno a carica e velocità per ottenere sempre lo stesso nella densità di corrente. Qui, però, stiamo affermando che è possibile *individuare* i responsabili del processo, cioè il singolo elettrone e la singola lacuna coinvolti, e dei siti specifici in cui avvengono i processi di generazione e ricombinazione. Tutto ciò non può verificarsi nei metalli, visto che gli elettroni sono “delocalizzati” nell’intero volume del sistema considerato.

È anche opportuno ricordare che questo modello trova una validazione sperimentale in misure basate sull’*effetto Hall*, che ben conoscete. Queste misure permettono anche di individuare diverse mobilità per i portatori n e p, essendo generalmente $\mu_p < \mu_n$ in un semiconduttore elementare a temperatura ambiente.

C. Equilibrio tra generazione e ricombinazione

A temperatura diversa da zero, il processo di *generazione* degli elettroni e la contestuale formazione di lacune avvengono continuamente. Altrettanto continuamente si verifica il processo, in qualche modo inverso, di *ricombinazione*. Mediando spazialmente e/o temporalmente, si potranno vedere questi due processi come i due versi di svolgimento di una sorta di “reazione chimica” attivata termicamente.

All’equilibrio dovrà valere la cosiddetta *legge di azione di massa*, che stabilisce $n_p \times n_n = \kappa(T)$, con n_p e n_n densità di lacune ed elettroni, rispettivamente, e $\kappa(T)$ costante specifica per il materiale (fortemente dipendente, però, dalla temperatura): nel silicio a temperatura ambiente si ha tipicamente $\kappa(T_{amb}) \sim 10^{20}$ (at/cm³)².

La legge di azione di massa ha alcune implicazioni pratiche importanti: (i) per un semiconduttore intrinseco, la densità dei portatori di carica, che deve essere la stessa per l’uno e l’altro segno, cioè $n_n = n_p$, vale $n_p = n_n \sim 10^{10}$ at/cm³, per cui, come già affermato, anche nel silicio intrinseco ci sono portatori di carica liberi; (ii) la densità di portatori di carica è parecchi ordini di grandezza (voi sapete quanto) minore che per un conduttore metallico; (iii) in ogni caso, i portatori di carica devono essere dei due segni, cioè elettroni e lacune devono convivere, altrimenti la legge di azione di massa non sarebbe rispettata.

L’aspetto rilevante che rende i semiconduttori di così grande interesse in elettronica è la possibilità di modificare, in maniera artificiale e controllata, n_n e n_p , cioè la densità di elettroni e lacune e, di conseguenza, di controllare le proprietà di conduzione. Questa bellissima opportunità, che nella pratica funziona solo con i semiconduttori, e soprattutto con il silicio, è alla base dell’aggettivo *bipolare* che spesso si associa al termine giunzione. In particolare, si attribuisce il simbolo “n” ai semiconduttori, o regioni di semiconduttore, in cui i portatori n sono in maggioranza e i p in minoranza, e il simbolo “p” alla situazione opposta.

III. DROGAGGIO

Si chiama drogaggio del semiconduttore l’operazione che porta alla *sostituzione*, all’interno del reticolo cristallino, di *alcuni* suoi atomi con atomi di un’altra specie. Per ovvie ragioni (dimensioni, “affinità” chimica, etc.), questa operazione non funziona con qualsiasi altra specie. Fortunatamente, però, facendo riferimento al silicio, essa va assai bene quando l’atomo di altra specie appartiene ai gruppi III o V della tavola periodica, per esempio è un atomo di boro (gruppo III) o di fosforo (gruppo V).

Esistono tante tecnologie, più o meno raffinate, per avere drogaggio: la più semplice prevede di riscaldare un *wafer* (una fetta cristallina) di silicio in un’atmosfera di composti contenenti, per esempio, boro o fosforo: questi elementi in determinate condizioni (in particolare a tem-

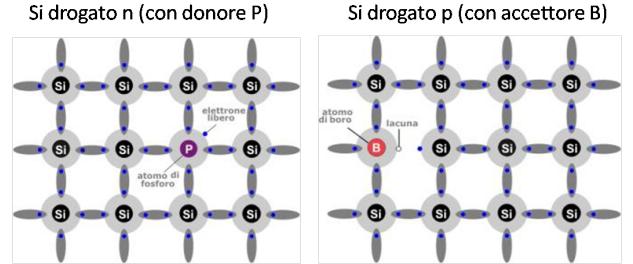


Figura 1. Visione schematica del drogaggio di Si con atomi donori (P) e accettori (B). Figura tratta da <http://www.marconi-galletti.it>.

perature di diverse centinaia di gradi centigradi) possono diffondere all’interno del wafer e lì sostituire localmente atomi di silicio del reticolo cristallino. Questa tecnologia è molto poco accurata nel determinare profili e densità di drogaggio: ne esistono tante ben più raffinate, e tuttora la messa a punto di tecniche ancora più avanzate è un importante argomento di ricerca in scienza dei materiali.

Supponiamo allora di aver sostituito alcuni atomi di silicio con atomi, per esempio, di fosforo: questi hanno la possibilità di coordinarsi con cinque atomi vicini, ma gli atomi vicini sono di silicio, che invece può coordinare solo con quattro atomi vicini. Un elettrone del fosforo non può formare il legame che vorrebbe, diventando così disponibile per il trasporto, cioè diventando di fatto un *portatore di carica negativa*, cioè di tipo n. Un drogante che conduce a questo risultato si chiama *donore*, dato che in un certo senso dona un proprio elettrone alla causa della conduzione.

Se invece considerassimo il boro come drogante, allora, dato che il boro può coordinare con solo tre atomi vicini, avremmo una situazione diametralmente opposta: invece di avere un elettrone in più che può partecipare alla conduzione, come nel caso del donore, ne avremmo uno in meno. Il drogante, in questo caso, si chiama *accettore* e il processo conduce alla realizzazione di una lacuna, cioè di un *portatore di carica positiva*, cioè di tipo p.

La Fig. 1 mostra una visione molto schematica di quello che avviene all’interno del cristallo di silicio nei due casi.

A. Silicio p e silicio n

Dunque in un pezzo di silicio (o di qualsiasi altro semiconduttore che possa essere drogato) si possono creare, artificialmente e in maniera controllata, densità di portatori di carica di un tipo o dell’altro diverse da quelle che si hanno intrinsecamente: in sostanza si possono creare semiconduttori *di tipo p* o *di tipo n* a seconda che essi siano drogati con accettori o con donori.

Alcune precisazioni prima di andare oltre:

- un pezzo di semiconduttore drogato è *globalmente neutro*: il drogaggio prevede di aggiungere atomi

(neutri) a un materiale semiconduttore che in origine è neutro, per cui non possono comparire in alcun modo eccessi di carica, né di un segno, né dell'altro. Notate infatti che, per esempio nel caso di un donore, l'elettrone in più disponibile per la conduzione lascia una carica positiva in più nello ione che costituisce il reticolo cristallino.

- Oltre ai portatori di carica libera creati attraverso il droggaggio, che si chiamano *maggioritari*, esistono sempre anche portatori di carica di segno opposto, detti *minoritari*, nello stesso materiale, necessari se non altro per soddisfare la legge di azione di massa.
- Densità atomiche tipiche per i droganti nel silicio possono variare tra circa 10^{11} e circa 10^{19} at/cm³. Tenete conto che la densità atomica del silicio è $\approx 5 \times 10^{22}$ at/cm³, per cui in ogni caso solo alcuni atomi di silicio del cristallo vengono sostituiti dal drogante. Dal punto di vista dei simboli, materiali fortemente drogati, per esempio di tipo p, si indicano spesso con p⁺ o addirittura p⁺⁺; analogamente, se di tipo n, con n⁺ o addirittura n⁺⁺. Un aspetto di dettaglio che vale la pena ricordare è che la legge di azione di massa, nella sua semplice formulazione che abbiamo dato prima, non vale per materiali fortemente drogati, detti “degeneri”.
- Anche per i semiconduttori drogati la meccanica quantistica stabilisce delle regole di carattere energetico relative, per esempio, alla promozione allo stato di conduzione dell'elettrone “spaiato” del donore. Continuiamo a disinteressarci di questi aspetti perché non strettamente necessari nel nostro modello, e anche perché la loro trattazione richiederebbe conoscenze di meccanica quantistica che non abbiamo.

IV. GIUNZIONE P-N

A questo punto abbiamo stabilito le premesse necessarie per trattare la giunzione di nostro interesse, che si realizza quando c’è un’interfaccia tra un pezzo di semiconduttore di tipo p e uno di tipo n. Propriamente parlando, il luogo geometrico dell’interfaccia si chiama *giunzione metallurgica*, dato che il termine giunzione tout-court si riferisce normalmente a tutto quello che succede anche “attorno” (o attraverso) l’interfaccia. Inoltre, se il materiale, droggaggio a parte, è lo stesso, per esempio silicio da ambo i lati, il termine corretto per descrivere il sistema è *omogiuinzione*, altrimenti si parla di *eterogiuinzione*. Qui faremo riferimento specifico alle omogiuinzioni di silicio.

Il sistema considerato ha la rappresentazione molto schematica di Fig. 2(a): dal punto di vista tecnologico, la giunzione metallurgica non può essere una reale discontinuità nel semiconduttore, nel senso che, per motivi che hanno a che fare con il comportamento delle superfici e in

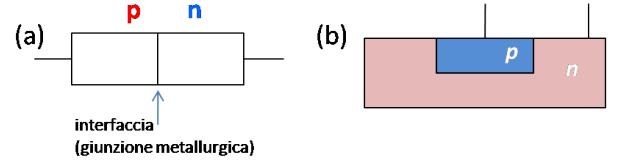


Figura 2. Visione molto schematica di una giunzione p-n (a) e sua rappresentazione un po’ più realistica e adeguata alle tecnologie attuali (b); i tratti lineari che escono dalle regioni p e n indicano dei fili di collegamento, necessari per il funzionamento del diodo costituito dalla giunzione.

particolare la loro ossidazione, è impensabile e fortemente sconsigliato produrre separatamente le due parti della giunzione e poi “metterli a contatto”. È infatti molto più efficace dal punto di vista delle proprietà elettriche (e anche della quantità di tecnologia coinvolta) partire da un unico pezzo di semiconduttore intrinseco e quindi drogare in modo diverso una parte rispetto all’altra. Inoltre, sempre per motivi tecnologici, la geometria della giunzione assume forme normalmente molto diverse rispetto a quella, a bacchetta, o lineare, di Fig. 2(a), e spesso somiglia di più allo schema di Fig. 2(b), che è topologicamente molto più compatibile con le tecnologie (“planari”) attualmente in uso.

Nonostante questa necessaria premessa, per scopi puramente didattici immagineremo di avere un pezzo di silicio drogato p e un pezzo drogato n, ovviamente entrambi neutri e inizialmente separati, e di metterli a contatto tra di loro.

A. Diffusione ed equilibrio

Per un banale motivo di *diffusione*, cioè per il semplice fatto che c’è un gradiente di densità dei portatori di carica liberi, dato che da una parte (p) c’è maggioranza di lacune e dall’altra (n) maggioranza di elettroni, si avrà immediatamente dopo il contatto uno spostamento di cariche: le lacune tenderanno a diffondere nel materiale drogato n, gli elettroni nel materiale drogato p.

Seguiamo il destino di una delle due specie, per esempio di un portatore n (un elettrone) che supera l’interfaccia per andare nella regione p: qui trova un’elevata densità di lacune e quindi ha un’elevata probabilità di ricombinarsi. Quando ricombinato, l’elettrone “prende casa” nel sito della lacuna, e quindi diventa, da libero quale era, una carica fissa che ha segno negativo e si stanzia nella regione p. Un processo simmetricamente opposto si verifica a carico delle lacune che in origine si trovavano nella regione p: si forma quindi una *separazione di carica fissa* che dà luogo alla formazione di un campo elettrico, diretto dalla regione n alla p, a cui si dà il nome di *campo di built-in*.

Il processo di diffusione attraverso l’interfaccia (segue la legge di *Fick*, come tutte le diffusioni dovute a gradiente di densità), a cui è associata una corrente di diffusione, procede fino al rapido raggiungimento di una condizione

di equilibrio. Esso è dovuto al campo elettrico di built-in che si oppone all'ulteriore migrazione di carica. All'equilibrio non c'è più migrazione di carica, e quindi la corrente di migrazione si annulla. All'equilibrio si crea una regione, attorno, o attraverso, l'interfaccia, *priva di cariche libere*: questa regione si chiama *regione di svuotamento* (di *depletion*), o anche *di carica spaziale* (di *space charge*). I termini sono evocativi: la regione è svuotata di cariche libere, ma in essa è presente carica fissa, di segno opposto a quella del droggaggio. Come troveremo un po' meglio nella prossima sezione, lo spessore della regione di svuotamento è normalmente molto ridotto.

Osservazione ovvia, ma importante: nella realtà microscopica l'equilibrio è *dinamico* (o cinetico), cioè il moto di diffusione, per esempio degli elettroni dalla regione n a quella p, prosegue continuamente per tante buone ragioni (la temperatura diversa da zero, o altre fluttuazioni), ma esso è continuamente bilanciato dal movimento in verso opposto dei portatori di carica minoritari (elettroni nella regione p che stiamo considerando). Mediando spazialmente e/o temporalmente si può affermare di raggiungere una condizione di equilibrio macroscopico.

V. PROBLEMINO DI ELETROSTATICA

Risolviamo, facendo uso di pesanti approssimazioni e limitandoci a una descrizione unidimensionale, il problemino di elettrostatica che è associato alla situazione considerata. Normalmente questo problemino è un esercizio standard di elettrostatica, ma vale comunque la pena di soffermarsi brevemente.

Modelliamo il nostro sistema come indicato in Fig. 3(a): nella regione p supponiamo di avere un eccesso di carica *negativa fissa* (all'equilibrio carica libera non ci può più essere, come stabilito in precedenza) con densità ρ_p (supposta in questo modello uniforme), e in quella n di avere un eccesso di carica *positiva fissa* con densità ρ_n (supposta in questo modello uniforme). Se ripensiamo all'origine microscopica di questi eccessi di carica e teniamo conto del loro segno, possiamo facilmente affermare che $\rho_p = -en_p$ e $\rho_n = en_n$, con e carica elementare (positiva), e n_p e n_n densità rispettive degli accettori nella regione p e dei donori in quella n. Si tratta quindi di parametri di fabbricazione della giunzione ed è interessante notare che agendo sulla concentrazione dei droganti e sulla sua distribuzione spaziale è possibile ingegnerizzare in maniera molto specifica il comportamento della giunzione.

Osservate che il risultante grafico della ρ in funzione della coordinata x diretta lungo l'asse della giunzione, e tale che $x = 0$ corrisponde all'interfaccia, ha andamenti "ripidi" che sono frutto della semplificazione di modello: andamenti più realistici dovrebbero prevedere una forma di smussamento degli spigoli, cioè densità di carica non uniformi. Inoltre notate che non è necessariamente vero che $|\rho_p| = |\rho_n|$, dato che non è necessariamente vero che $n_p = n_n$: le due regioni della giunzione sono infatti diverse, cioè costruite in modo diverso, e in genere

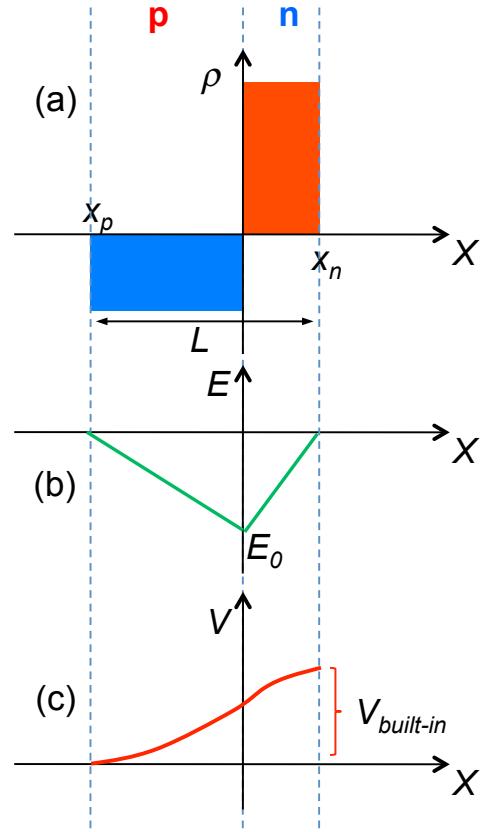


Figura 3. Visione schematica e qualitativa della densità di carica (a), del campo (b), e della differenza di potenziale (c) rappresentati in funzione della coordinata x per la giunzione p-n (p a sinistra, n a destra) considerata nel testo. Nel pannello (c) si è posto nullo il potenziale per $x \leq x_p$; il simbolo $V_{built-in}$ indica il potenziale di built-in.

hanno anche densità di drogante differenti. Per garantire la neutralità dell'intero sistema la quantità di carica complessiva deve essere nulla: se $n_p \neq n_n$, questo implica una regione di svuotamento, o, se preferite, giunzione tout-court, asimmetrica, estesa maggiormente nella regione in cui la densità di drogante è minore, ma tale che l'integrale sotteso alla curva ρ in funzione di x sia nullo.

L'equazione di Poisson per il campo E (non usiamo il segno di vettore perché esaminiamo un problema unidimensionale), scritta nella sola direzione x , cioè la $dE/dx = \rho/\epsilon$, recita

$$\frac{dE}{dx} = -\frac{en_p}{\epsilon} \text{ regione p } (x < 0) \quad (1)$$

$$\frac{dE}{dx} = \frac{en_n}{\epsilon} \text{ regione n } (x > 0) . \quad (2)$$

Il nostro modello prevede che la costante dielettrica, indicata con ϵ , sia uniforme in tutta la giunzione e pone $\rho = 0$ per $x = 0$, cioè sul piano di interfaccia non c'è alcun eccesso di carica.

Integrando le Eqs. 1, 2 tra gli estremi $x_p, 0$ per $x < 0$, e $0, x_n$ per $x > 0$, e supponendo ragionevolmente che il

campo sia continuo all'interfaccia ($E_0^- = E_0^+ = E_0$ per $x \rightarrow 0^-$ e $x \rightarrow 0^+$), si ottiene

$$-E_{x=x_p} + E_0 = \frac{en_p}{\epsilon} x_p \quad (3)$$

$$E_{x=x_n} - E_0 = \frac{en_n}{\epsilon} x_n . \quad (4)$$

Nelle condizioni che stiamo supponendo (giunzione all'equilibrio e *non polarizzata*, termine che chiariremo nel seguito), è ragionevole porre le condizioni al contorno (sul campo) $E_{x=x_p} = E_{x=x_n} = 0$, da cui

$$E_0 = \frac{en_p}{\epsilon} x_p = -\frac{en_n}{\epsilon} x_n , \quad (5)$$

che, notate, indica un campo negativo e che dà, come conseguenza, la condizione di neutralità $|n_p x_p| = |n_n x_n|$ già incontrata prima.

Nell'ambito di questo modello, lo spessore $L = (x_n - x_p)$ della regione di svuotamento è legato al valore di E_0 attraverso la densità di droggaggio nelle due regioni e la costante dielettrica, entrambi specifici del sistema considerato. Si ottiene infatti

$$L = -\frac{\epsilon E_0}{e} \left(\frac{1}{n_n} + \frac{1}{n_p} \right) ; \quad (6)$$

dunque aumentando la densità dei droganti si ottiene una riduzione dello spessore della regione di svuotamento.

A. Campo e potenziale

Integrando le Eqs. 1, 2 tra x generico e zero in tutte e due le regioni, e tenendo conto di quanto trovato, si ottiene con pochi passaggi

$$E(x) = \frac{en_p}{\epsilon} (x_p - x) \text{ regione p } (x < 0) \quad (7)$$

$$E(x) = \frac{en_n}{\epsilon} (x - x_n) \text{ regione n } (x > 0) . \quad (8)$$

Si vede allora che in questo modello l'intensità del campo elettrico è una funzione lineare della coordinata x all'interno delle singole regioni e segue l'andamento di Fig. 3(b).

Determiniamo ora la differenza di potenziale ΔV_{junct} tra i due estremi della regione di svuotamento, ovvero attraverso la giunzione. Si ha

$$\Delta V_{junct} = - \int_{x_p}^{x_n} E(x) dx = -\frac{e}{\epsilon} \left[\int_{x_p}^0 n_p (x_p - x) dx + \int_0^{x_n} n_n (x - x_n) dx \right] . \quad (9)$$

$$+ \int_0^{x_n} n_n (x - x_n) dx . \quad (10)$$

Svolgendo i calcoli, si ottiene

$$\Delta V_{junct} = \frac{e}{2\epsilon} (n_p x_p^2 + n_n x_n^2) . \quad (11)$$

Poichè nelle regioni del sistema al di fuori della regione di svuotamento abbiamo ragionevolmente supposto campo

nullo, questa è anche la d.d.p. che esiste tra l'estremo "sinistro" (regione p) e quello "destro" (regione n) dell'intero sistema.

In condizioni di equilibrio, c'è allora una differenza di potenziale positiva se si passa dalla regione drogata p alla regione drogata n. Si dice allora che esiste una *barriera di potenziale* di altezza ΔV_{junct} , chiamata in genere *potenziale di built-in* ($V_{built-in} = \Delta V_{junct}$), dato che essa ha origine dal campo di built-in. $V_{built-in}$ dipende soprattutto dal materiale e, nel silicio, si ha $V_{built-in} \simeq 0.7$ V.

Naturalmente il *potenziale* elettrico nella regione di svuotamento dipende dalla coordinata x , come si può facilmente determinare integrando l'Eq. 9 tra x_p e x generico. Avendo scelto (arbitrariamente) nullo il potenziale per $x \leq x_n$, si ottiene con un po' di passaggi la seguente espressione per $V(x)$:

$$V(x) = \frac{e}{2\epsilon} n_p (x_p - x)^2 \text{ regione p } (x < 0) \quad (12)$$

$$V(x) = \frac{e}{2\epsilon} [(n_p x_p^2 + n_n x_n^2) - n_n (x - x_n)^2] \quad (13)$$

$$\text{regione n } (x < 0) , \quad (14)$$

Questo potenziale è di norma sempre crescente: un andamento è rappresentato in Fig. 3(c).

Infine, avendo stabilito che per il silicio tipicamente $\Delta V_{junct} = V_{built-in} \simeq 0.7$ V, possiamo stimare nell'ambito del nostro modello lo spessore L della regione di svuotamento. Nel caso, semplice, $n_p = n_n = n$, e dunque $x_p = x_n$, si ha infatti da Eq. 11 $L = \sqrt{2\epsilon V_{built-in}/(en)}$; il valore numerico, calcolato usando $\epsilon_{r,Si} = 11.7$, fornisce per esempio una stima $L \simeq 300$ nm per $n = 10^{16}$ at/cm³, e $L \simeq 30$ nm per $n = 10^{18}$ at/cm³, che costituiscono una valutazione molto grossolana, vista la semplicità del modello, dello spessore della regione di svuotamento, ma sono in ragionevole accordo con previsioni più accurate.

VI. DIODO

Un diodo a giunzione p-n è sostanzialmente una giunzione di quelle esaminate finora con due elettrodi conduttori alla fine delle parti di semiconduttore p e n, e due fili, o terminali, o reofori, che ne escono. Il diodo è un componente circuitale a due terminali evidentemente non intercambiabili fra di loro. Il simbolo, rappresentato in Fig. 4, è infatti asimmetrico (la freccia, come capiremo fra breve, indica la direzione di passaggio della corrente): la parte costituita dal semiconduttore p si chiama *anodo* (A), quella n *catodo* (K). Il terminale corrispondente al catodo è normalmente indicato sull'involucro esterno (in genere cilindrico) con una fascetta, ma ci sono anche tante altre convenzioni, dipendenti anche dalla forma dell'involucro.

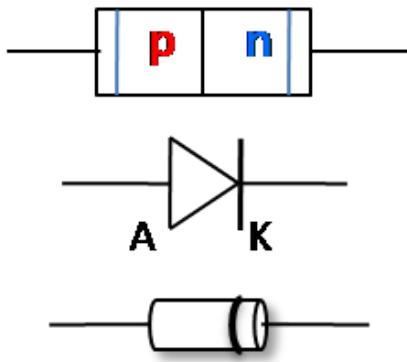


Figura 4. Rappresentazione della giunzione p-n, simbolo del diodo, ed esempio di diodo munito di involucro esterno, con una fascetta a indicare il catodo.

A. Diodo non polarizzato e polarizzazione

Nell'uso circuituale, a un diodo può essere applicata la d.d.p. di un generatore esterno; si può quindi formare un circuito attraverso il quale può passare una certa corrente. I terminali del diodo servono proprio per queste funzioni: quando essi non sono collegati a nulla, ovvero non c'è alcuna d.d.p. "esterna" applicata, la giunzione assume le condizioni di equilibrio di cui alla sezione precedente. Si dice in questo caso che il diodo è *non polarizzato*, altrimenti si dice che si è applicata al diodo una certa *polarizzazione*.

In assenza di polarizzazione, il campo elettrico di built-in è orientato dalla regione n a quella p, cioè dal catodo all'anodo. Esso è in grado di bloccare (all'equilibrio) la diffusione dei portatori di carica, in accordo con quanto determinato in precedenza. La situazione è schematicamente rappresentata in Fig. 5(a), che mostra anche l'andamento qualitativo della differenza di potenziale tra catodo e anodo, in cui si nota la presenza di una barriera di potenziale. In queste condizioni ovviamente non c'è passaggio di corrente attraverso la giunzione.

Osserviamo che la configurazione di carica che abbiamo supposto produce effetti sostanzialmente simili a quelli dovuti alle cariche che si trovano sulle armature di un condensatore piano parallelo, in particolare la formazione di un campo elettrico omogeneo. Fate attenzione, però, al fatto che nella regione di svuotamento esistono eccessi di carica (comunque fissa e non mobile), cosa che si suppone impossibile in un condensatore convenzionale.

B. Polarizzazione inversa

Si dice che il diodo è posto a *polarizzazione inversa* quando ai suoi terminali è applicata una d.d.p. esterna, cioè prodotta dall'esterno, orientata in modo che il polo negativo sia collegato al terminale della regione p, cioè all'anodo, e quello positivo al terminale della regione n, cioè al catodo.

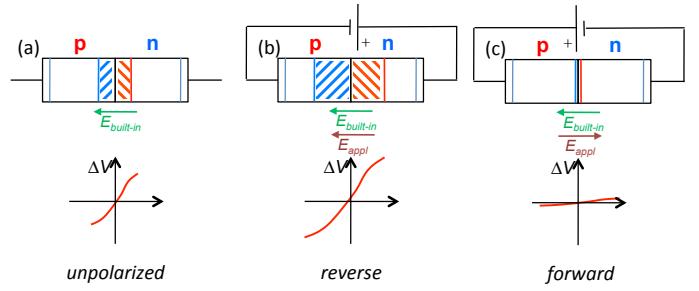


Figura 5. Rappresentazione schematica dell'eventuale collegamento con il generatore di d.d.p. esterno (d.d.p. prodotta esternamente e "applicata" al diodo), della distribuzione di carica nella giunzione e della barriera di potenziale ΔV tra i terminali nei casi di diodo non polarizzato (a), polarizzato inversamente (b) e direttamente (c). I vettori $\vec{E}_{\text{built-in}}$ e \vec{E}_{appl} indicano rispettivamente le direzioni e i versi del campo di built-in e dell'eventuale campo legato all'applicazione della d.d.p. esterna. Le regioni tratteggiate rappresentano le zone in cui si trovano cariche fisse (regioni di *carica spaziale*), con segno identificato dal colore (blu e rosso rispettivamente per negativo e positivo).

L'applicazione di questa d.d.p. produce, o corrisponde a, un campo elettrico nella zona compresa tra gli elettrodi, che, in prima approssimazione, possiamo considerare uniforme. Come mostrato in Fig. 5(b), questo campo è parallelo a quello generato per built-in nella giunzione. Senza entrare in aspetti modellistici e ripetere qualcosa di simile ai calcoli necessari per la soluzione del problema affrontato sopra, ci limitiamo a considerare l'effetto di questo campo sulla barriera di potenziale esistente tra i due lati della giunzione, cioè tra i due terminali del diodo: si conclude facilmente che, in queste condizioni, l'altezza della barriera di potenziale viene aumentata, dato che ad essa contribuisce (con lo stesso segno) la d.d.p. di polarizzazione.

C'è poi, a livello qualitativo, un altro effetto notevole: nella descrizione di modello che stiamo adottando, l'applicazione della d.d.p. in polarizzazione inversa *aumenta lo spessore della regione di svuotamento*. Il campo prodotto dalla d.d.p. esterna tende infatti a "richiamare" verso i rispettivi elettrodi le lacune nella regione p e gli elettroni nella regione n. Queste cariche sono quelle coinvolte nei processi di ricombinazione che seguono la diffusione: per intenderci, gli elettroni che diffondono nella regione p, quando la polarizzazione è inversa, incontrano comparativamente meno (una densità minore di) lacune, poiché queste si sono spostate verso l'anodo, cioè lontano dall'interfaccia. Pertanto questi elettroni possono penetrare, per diffusione, più in profondità nella regione p prima di ricombinare. Alla fine, la regione di svuotamento risulta più estesa, ovvero più spessa, come indicato schematicamente in Fig. 5(b). In un certo senso, il campo elettrico applicato alla giunzione nel caso di polarizzazione inversa tende a "favorire" gli effetti della diffusione responsabili per la creazione della regione di

svuotamento.

Osservate un interessante effetto collaterale nel diodo polarizzato inversamente. Come già affermato, la presenza del campo elettrico di built-in nella regione di svuotamento fa somigliare la configurazione considerata a quella di un condensatore ad armature piane e parallele. Polarizzando inversamente si aumenta l'equivalente dello spessore di dielettrico tra le armature del condensatore, cioè diminuisce la *capacità del condensatore costituito dalla giunzione*. Pur nella sua semplicità, questa affermazione trova riscontro nel funzionamento dei cosiddetti *diodi varicap*, usati proprio come condensatori variabili comandati dalla d.d.p. di polarizzazione inversa.

In condizioni di polarizzazione inversa, a causa della presenza di una barriera di potenziale, ci si aspetta che non scorra corrente attraverso la giunzione, e dunque attraverso il circuito formato da diodo e generatore: il diodo si trova quindi *in interdizione*. In realtà, però, si deve assumere la possibilità che ci sia una corrente, molto debole, in grado di scorrere nel circuito, la *corrente di saturazione inversa*. Questa corrente non potrà essere fatta di portatori maggioritari. Infatti un portatore maggioritario che si sposta da anodo a catodo è una lacuna nella regione p: essa vede una *differenza di energia potenziale* $\Delta U_{magg} = e\Delta V_{junct}$, che è positiva e che quindi classicamente non può essere superata dalla lacuna. Però, come abbiamo ampiamente discusso, in tutte e due le parti della giunzione esiste una densità piccola, ma non nulla, di portatori *minoritari*; essi sono elettroni, carichi negativamente, nella regione p e quindi vedono una differenza di energia potenziale $\Delta U_{min} = -e\Delta V_{junct}$. Essa è negativa, e quindi può essere superata dalle cariche minoritarie che quindi sono messe in movimento dal generatore dando luogo alla debole corrente di saturazione inversa. Il termine saturazione suggerisce che essa non dipende dalla d.d.p. applicata (purché del segno giusto). La corrente di saturazione inversa è generalmente molto piccola, e spesso con effetti pratici trascurabili; come accenneremo in seguito, esiste però un ulteriore effetto che rende possibile avere una corrente inversa molto intensa per valori della d.d.p. generalmente elevati, cioè dell'ordine di diverse decine di V.

C. Polarizzazione diretta

La *polarizzazione diretta* si realizza collegando un generatore di d.d.p. esterna con il polo positivo al lato p (anodo) e quello negativo al lato n (catodo). È immediato intuire che gli effetti sono opposti a quelli discussi per la polarizzazione inversa.

Infatti il campo dovuto all'applicazione della d.d.p. esterna è in questo caso anti-parallelo rispetto a quello di built-in. Di conseguenza la barriera di potenziale tra i due lati della giunzione si abbassa, in maniera tanto più marcata quanto maggiore è la d.d.p. applicata al diodo in polarizzazione diretta, vedi Fig. 5(c). Al di sopra di un certo valore di questa d.d.p., grossolanamente individua-

ta nella *tensione di soglia*, V_{thr} , il diodo *entra in conduzione*, cioè lascia passare rilevanti intensità di corrente, dove per "rilevanti" si può intendere ben maggiori dell'intensità di saturazione inversa (accenneremo in seguito a valori tipici).

La tensione di soglia è, come facile intuire, dello stesso ordine di grandezza di $V_{built-in}$, anche se non esattamente corrispondente a questa a causa di diversi aspetti di dettaglio che non esaminiamo qui. Osservate che molto spesso si dice che la giunzione è *polarizzata direttamente* solo quando il diodo è entrato in conduzione: questo può creare confusione rispetto alle situazioni sperimentali di polarizzazione diretta *sotto soglia*, in cui la corrente che attraversa la giunzione e interessa il circuito è comparativamente piccola.

Due osservazioni sottili, ma rilevanti:

- ripetendo i ragionamenti presentati in precedenza per la giunzione polarizzata inversamente, è chiaro che si può giungere ad affermare che lo spessore della giunzione tende a diminuire quando la polarizzazione è diretta. Si potrebbe quindi azzardare che lo spessore diventi nullo, cioè, intuitivamente, non ci sia più giunzione, quando si arriva alla tensione di soglia. Questa conclusione non è corretta per vari motivi.
- Il motivo più evidente a supporto della necessità che esista sempre e comunque un certo spessore associato alla regione di svuotamento discende dal fatto che, come è evidente tenendo conto dei versi del campo elettrico prodotto dal generatore, le cariche coinvolte nel processo di conduzione in polarizzazione diretta sono sempre e solo *portatori maggioritari*, cioè lacune ed elettroni nelle regioni rispettivamente p e n. Facendo riferimento alla Fig. 5(c), in queste condizioni della carica positiva entra nella regione p dalla sinistra di figura, per cui ci sono delle lacune che viaggiano in questa regione verso l'interfaccia. Nella regione n, invece, ci sono degli elettroni che entrano dalla destra di figura e si muovono verso l'interfaccia. Dunque i portatori di carica coinvolti nel processo "cambiano segno" passando attraverso l'interfaccia. Questo può avvenire solo invocando la ricombinazione che, come affermato sopra, deve necessariamente avvenire nella regione di svuotamento. Quindi, per quanto di spessore ridotto, la regione di svuotamento deve esistere.

VII. MODELLO DI SHOCKLEY E CURVA CARATTERISTICA

La descrizione fatta finora permette di cogliere gli aspetti più eclatanti del funzionamento di un diodo a giunzione p-n: il passaggio di corrente attraverso questo componente dipende evidentemente dal segno della

d.d.p. applicata, e anche, nel caso di polarizzazione diretta, dal valore di tale d.d.p., che deve superare una certa soglia per avere un consistente passaggio di corrente.

È altrettanto evidente che quello della giunzione p-n è un ottimo esempio di *comportamento non ohmico*: infatti la legge di Ohm, che vale per esempio nel caso dei conduttori, e dunque dei resistori che ne fanno uso, prevede una dipendenza lineare tra corrente e tensione applicata, a prescindere dal segno.

La verifica del comportamento ohmico o non ohmico di un componente si esegue agevolmente tracciando la *curva caratteristica I-V* del componente stesso, cioè osservando l'andamento dell'intensità di corrente in funzione della d.d.p. applicata, segno compreso. Nel caso di componenti a due terminali (si chiamano qualche volta *bipoli*) non c'è alcuna ambiguità e si capisce benissimo di quale tensione e quale corrente si parli.

Una ragionevole descrizione analitica della curva caratteristica, almeno per determinati regimi di operazione, è data dal *modello di Shockley*, originariamente sviluppato per interpretare le prime giunzioni costruite basate su germanio invece che silicio. La fisica che sta dentro il modello di Shockley prevede un po' di meccanica quantistica e, soprattutto, una descrizione statistica dei portatori di carica che non possiamo, né vogliamo, affrontare qui. Ci limitiamo allora a scrivere l'equazione analitica della curva caratteristica ottenuta da questo modello:

$$I = I_0 \left[\exp \left(\frac{\Delta V}{\eta V_T} \right) - 1 \right]. \quad (15)$$

Nell'equazione, ΔV indica la d.d.p., o tensione, applicata al diodo, mentre I_0 coincide praticamente con l'intensità di corrente di saturazione inversa citata in precedenza. Il fattore che è al denominatore dell'esponenziale contiene un parametro "ingegneristico", η , che dipende dai dettagli costruttivi e soprattutto dal materiale della giunzione (vale uno per il germanio e circa 2 per un ordinario diodo al silicio), e la d.d.p. V_T . Questa è legata alla temperatura T a cui si trova la giunzione attraverso la seguente relazione:

$$eV_T = k_B T, \quad (16)$$

con k_B costante di Boltzmann. A temperatura ambiente ($T \simeq 300$ K), come già affermato il secondo membro dell'equazione vale $k_B T \simeq 1/40$ eV, valore da ricordare a vita, per cui si ha $V_T \simeq 26$ mV.

La specifica dipendenza esplicita dalla temperatura espressa dall'Eq. 15 ha a che fare con l'attivazione termica di processi di generazione e ricombinazione tipici dei semiconduttori. Osservate, però, che questa dipendenza, per la quale l'intensità di corrente a polarizzazione diretta diminuisce (e quindi la resistenza aumenta) con l'aumentare della temperatura, è affiancata dalla dipendenza con la temperatura della corrente di saturazione inversa I_0 . Infatti essa è ovviamente proporzionale alla densità dei portatori di carica minoritari, che, si pensi alla legge di azione di massa, è funzione della temperatura. Normalmente è proprio la dipendenza di I_0 da T a

conferire gli effetti più marcati nel comportamento di un diodo con la temperatura.

Il grafico della curva caratteristica I-V costruito analiticamente secondo l'Eq. 15 e mostrato come esempio in Fig. 6 separatamente per $\Delta V > 0$ e $\Delta V < 0$, ha le seguenti caratteristiche salienti:

- per $\Delta V < 0$ tende rapidamente al valore $-I_0$: per gli ordinari diodi al silicio in uso in laboratorio, per esempio il modello 1N914, si ha $I_0 \sim 1 - 10$ nA, che illustra bene il significato di "molto piccola" che abbiamo usato in precedenza per definire la corrente di saturazione inversa.
- La corrente è nulla per $\Delta V = 0$, ovviamente e in accordo con quanto affermato per la giunzione non polarizzata.
- Per $\Delta V > 0$ l'intensità di corrente si mantiene relativamente bassa fino a una sorta di *ginocchio*, al di sopra della quale il comportamento esponenziale diviene evidentissimo: in questo regime il termine esponenziale in Eq. 15 prevale grandemente sul termine -1 .
- Il ginocchio di cui sopra indica in modo chiaro l'esistenza di una tensione di soglia V_{thr} ; tuttavia, anche facendo uso del modello di Shockley, V_{thr} rimane indeterminato (si può solo affermare che, per un ordinario diodo al silicio, è tipicamente $V_{thr} \sim 0.45 - 0.65$ V). Infatti è evidente che la collocazione del ginocchio nel grafico I-V dipende dalla scala del grafico stesso. Non esiste una definizione univoca per V_{thr} , che qualche volta viene fatto corrispondere al valore di tensione in cui I raggiunge l'1% del valore di intensità di corrente massima tollerata dal componente, come dichiarata dal costruttore. Per i diodi ordinari in uso in laboratorio tale intensità massima è dell'ordine di 300 mA, per cui dall'Eq. 15, supponendo $I_0 = 6$ nA, $\eta = 2$ e $T = 300$ K, si ricava $V_{thr} \sim 0.66$ V.

- Un'ovvia conseguenza del comportamento fortemente non lineare del diodo, e, in generale, di tutte le giunzioni p-n, è che la *resistenza effettiva* della giunzione *dipende marcatamente* dalle condizioni di lavoro, in particolare dalla corrente che attraversa il componente. Infatti la resistenza è definita dal rapporto tra d.d.p. ΔV e intensità di corrente I , che è rilevante per basse correnti (o basse tensioni), mentre tende a divenire trascurabile sopra soglia.

A. Ulteriori precisazioni sulla caratteristica I-V

Benché l'Eq. 15 descriva piuttosto bene la maggior parte delle situazioni sperimentali in cui si fa uso di diodi a giunzione p-n, essa non è accurata nel prevedere il comportamento in determinate condizioni, in particolare

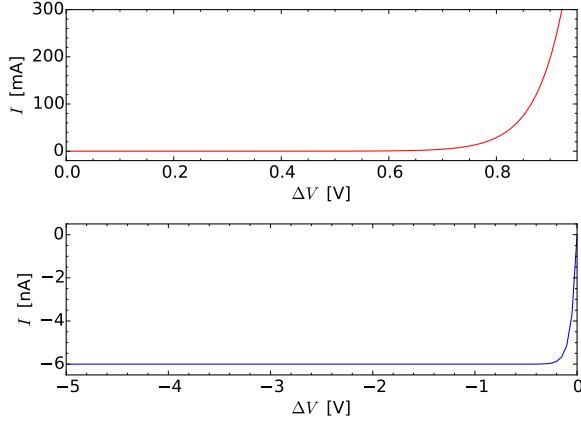


Figura 6. Grafico dell'Eq. 15 calcolato per $I_0 = 6$ nA, $\eta = 2$, $V_T = 26$ mV: il pannello superiore si riferisce a $\Delta V > 0$ (polarizzazione diretta), con limite, scelto per chiarezza, a $\Delta V = 0.95$ V; quello inferiore a $\Delta V < 0$ (polarizzazione inversa), con limite a $\Delta V = -5$ V. Notate le diverse scale verticali dei due grafici (mA e nA per polarizzazione rispettivamente diretta e inversa).

per giunzioni fortemente polarizzate negativamente e per elevate correnti in polarizzazione diretta.

1. Breakdown ed effetto Zener

Per ΔV molto negativo, cioè $\Delta V < V_{br}$ (con $V_{br} \sim -80 \text{--} -100$ V, nei diodi al silicio di uso comune) i portatori minoritari coinvolti nella conduzione inversa possono raggiungere velocità di deriva classiche molto elevate. Infatti il campo elettrico dovuto alla d.d.p. applicata può diventare molto intenso, vista la piccola lunghezza complessiva del dispositivo, tipicamente ben sotto 1 mm (ricordate che il campo, approssimando il sistema come un condensatore a simmetria piana, dipende inversamente dalla distanza tra gli elettrodi, cioè la “lunghezza” di cui sopra). In queste condizioni si può avere un fenomeno di *breakdown* dovuto alla ionizzazione degli atomi del reticolo cristallino per collisione con i portatori di carica, la cui densità aumenta in un processo a valanga che, di fatto, annulla la resistenza della giunzione.

Osservate che questo fenomeno non può verificarsi in polarizzazione diretta, dove la d.d.p. applicata può assumere al massimo valori di poco superiori a V_{thr} , come discuteremo nel seguito. Il fenomeno di breakdown è spesso distruttivo negli ordinari diodi al silicio, però esso può essere ingegnerizzato in modo da avvenire a tensioni relativamente basse e ben controllate, dell'ordine di pochi V, senza dare luogo a danneggiamento. I dispositivi che sfruttano tale ingegnerizzazione si chiamano *diodi Zener*, e l'ingegnerizzazione che stiamo ipotizzando consiste, in sostanza, nell'innescare il processo di breakdown grazie a un meccanismo di trasporto di carica puramente quantistico, legato all'*effetto tunnel*. Il fatto di opporre

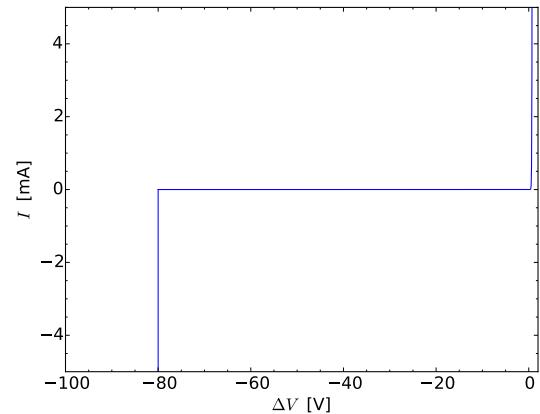


Figura 7. Rappresentazione semi-qualitativa del comportamento della curva I-V per polarizzazione inversa fino a valori dell'ordine di V_{br} , qui scelto $V_{br} = -80$ V, come tipicamente riscontrato negli ordinari diodi al silicio.

una resistenza effettiva praticamente nulla per un preciso valore di d.d.p. applicata è utile per costruire dei *riferimenti di tensione*, in cui, come rappresentato in Fig. 7, la d.d.p. rimane costante a prescindere dalla corrente che passa nel dispositivo.

2. Resistenza interna dei semiconduttori

Un'altra modifica all'Eq. 15 può diventare rilevante quando, in condizioni di polarizzazione diretta, il diodo si trova a essere interessato da valori di I relativamente alti, per esempio a partire dalla decina di mA nei diodi al silicio di uso comune. In queste condizioni si osserva una progressiva deviazione dall'andamento esponenziale previsto dal modello di Shockley, che tende a evolvere verso un andamento lineare (ohmico). Questa è la manifestazione della presenza di elementi resistivi, le parti “più esterne” dei semiconduttori p e n, in serie alla giunzione. Se per la giunzione la resistenza effettiva assume un valore molto piccolo, dovuto al fatto che si suppone di lavorare a $\Delta V \gg V_{thr}$, le parti esterne del dispositivo, quelle al di fuori della regione di svuotamento (elettrodi e fili compresi), continuano a manifestare un comportamento resistivo, cioè ohmico, dovuto alla mobilità finita dei portatori di carica (maggioritari) nel semiconduttore, oltre che a qualsiasi altra resistenza di contatto. Ovviamamente la resistenza associata a questi effetti è bassa, al punto che, di norma, essa diventa praticamente trascurabile per basse correnti, dove invece prevale la resistenza della giunzione.

Per evidenziare gli effetti della resistenza interna, l'Eq. 15 è stata invertita per $\Delta V > V_{thr}$, dove $I \sim I_0 \exp(\Delta V / (\eta V_T))$, in modo da ottenere la d.d.p. in funzione della corrente I che attraversa il diodo (notate che, per i valori di d.d.p. considerati, il termine -1 è trascurabile rispetto all'esponenziale). Quindi si è aggiunto un

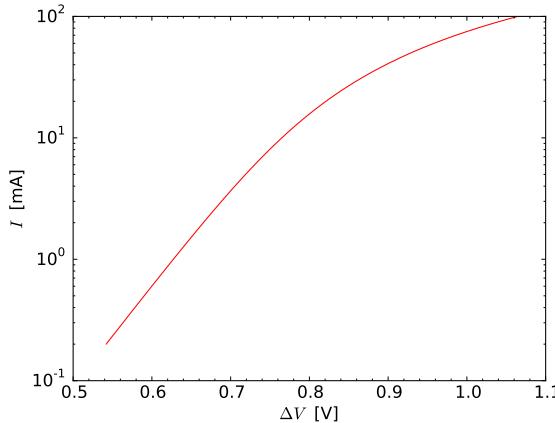


Figura 8. Grafico I-V calcolato come discusso nel testo tenendo conto della caduta di potenziale sulla resistenza dei materiali semiconduttori e dei contatti, qui assunta $R_{int} = 2$ ohm. Si noti la scala semilogaritmica: la deviazione dall'andamento lineare evidenzia l'effetto di questa sorta di resistenza interna.

termine $R_{int}I$ che tiene conto della caduta di potenziale sulla resistenza “interna”, ottenendo

$$\Delta V = \eta V_T \ln(I/I_0) + R_{int}I . \quad (17)$$

Infine, si è rappresentata in scala semilogaritmica la funzione ponendo, per analogia con i grafici precedenti, ΔV sull’asse orizzontale e I su quello verticale. Il grafico in scala semilogaritmica è atteso avere un andamento lineare se il termine R_{int} può essere considerato trascurabile: dunque ogni variazione rispetto all’andamento lineare è segno che la resistenza interna gioca un ruolo non trascurabile.

La Fig. 8 mostra il risultato del calcolo eseguito supponendo gli stessi parametri di Fig. 6 e ponendo $R_{int} = 2$ ohm (un valore piuttosto ragionevole per gli ordinari diodi al silicio): si riscontra un’evidente deviazione dal comportamento lineare (in carta semilogaritmica) a partire da $I \lesssim 10$ mA. Questa deviazione è anche ben visibile nelle curve I-V riportate nei datasheet, dove spesso si impiega una rappresentazione semilogaritmica.

Benché gli effetti di R_{int} siano generalmente trascurabili a bassi valori di corrente, la presenza di una resistenza al di fuori della giunzione, e in serie con questa, può facilmente influenzare la determinazione sperimentale della tensione di soglia V_{thr} in dispositivi reali. Infatti su questa resistenza si determina una caduta di potenziale, che è normalmente responsabile per l’aumento apparente della tensione di soglia fino a valori che spesso superano 0.7 V.

VIII. RESISTENZA DELLA GIUNZIONE P-N

Sulla base di quanto abbiamo ripetutamente affermato a proposito dell’evidente comportamento non-ohmico

del diodo a giunzione p-n, parlare di resistenza del componente può suonare fuori luogo. Infatti la resistenza di un componente ohmico è *definita* come $R = \Delta V/I$, relazione che implicitamente presuppone che la differenza di potenziale ΔV ai capi del componente stesso dipenda linearmente dalla corrente I che vi fluisce attraverso. Come visto, questa dipendenza lineare non è proprio verificata nel diodo, e quindi per tale componente non si può individuare una resistenza R costante, cioè indipendente dalle condizioni di operazione.

Tuttavia fa spesso comodo utilizzare ancora il concetto di resistenza, o *resistenza efficace*, anche nel caso di un diodo, o, più in generale, di una giunzione, cosa ragionevole fatte salve, però, alcune precisazioni, come quelle che discuteremo qui nel seguito.

A. Retta di carico e soluzione grafica

Supponiamo di avere un circuito semplicissimo, costituito da un generatore di d.d.p. *reale*, con resistenza interna r e tensione a circuito aperto V_0 , a cui è collegato un diodo; supponiamo anche che il collegamento sia realizzato in modo tale che il diodo si trovi in polarizzazione diretta. Chiediamoci quanto vale la corrente I che fluisce nel diodo e quindi nel semplicissimo circuito considerato.

Se il diodo fosse un componente ohmico la risposta sarebbe immediata, potendo usare direttamente la legge di Ohm (ovvero le regoline dei partitori di tensione). Qui, invece, dobbiamo tenere conto del legame che esiste tra la d.d.p. ai capi del diodo, che continuiamo a chiamare ΔV , e la corrente che fluisce nel circuito, che continuiamo a chiamare I . In sostanza, tenendo conto della caduta di potenziale ai capi della resistenza interna del generatore, possiamo scrivere questo sistema di due equazioni:

$$V_0 = rI + \Delta V \quad (18)$$

$$I = I_0 \left[\exp\left(\frac{\Delta V}{\eta V_T}\right) - 1 \right] , \quad (19)$$

dove la prima è sostanzialmente l’applicazione della legge di Ohm alla resistenza interna del generatore, mentre nella seconda riconosciamo l’equazione di Shockley.

La soluzione del sistema, in particolare la determinazione di I essendo noti tutti gli altri parametri in gioco, richiede tipicamente un approccio numerico, data la dipendenza non lineare stabilita dal modello di Shockley. Tradizionalmente (dai tempi in cui i computer non esistevano o non erano abbastanza diffusi e le curve caratteristiche dei diodi si trovavano facilmente nei manuali) la soluzione si può ottenere in modo grafico. Basta infatti trovare l’intersezione tra la curva caratteristica I-V del diodo che si sta utilizzando con la curva, detta *retta di carico*

$$I = \frac{V_0}{r} - \frac{\Delta V}{r} \quad (20)$$

disegnata sullo stesso piano. Questa curva, che è evidentemente una retta, ha *pendenza negativa* e intercetta gli

assi nei punti $V = V_0$ e $I = V_0/r$, per cui è perfettamente determinata essendo noti V_0 e r . L'intersezione fra la retta di carico e la curva caratteristica del diodo determina il cosiddetto *punto di lavoro*, permettendo in particolare di stabilire la *corrente di lavoro* I_q che rappresenta la corrente che passa nel circuito considerato.

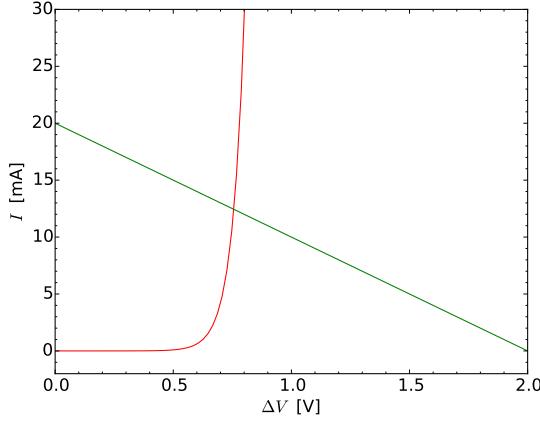


Figura 9. Esempio di soluzione grafica per la determinazione del punto di lavoro di un diodo, modellato con gli stessi parametri di Fig. 15, collegato a un generatore reale di d.d.p. $V_0 = 2$ V (a circuito aperto) e resistenza interna $r = 100$ ohm. La linea verde rappresenta la retta di carico e il punto di intersezione con la curva caratteristica stabilisce la corrente di lavoro $I_q \simeq 12$ mA del diodo.

La Fig. 9 mostra un esempio di individuazione grafica: la curva caratteristica del diodo è stata calcolata usando gli stessi parametri di Fig. 6, mentre per la retta di carico, disegnata in verde, si è supposto $V_0 = 2$ V e $r = 100$ ohm, per cui essa intercetta gli assi in $\Delta V = 2$ V e $I = 20$ mA. Il punto di incrocio fra retta di carico e curva del diodo determina la corrente di lavoro $I_q \simeq 12$ mA. Notate che, a causa dell'andamento esponenziale della curva caratteristica, la tensione di lavoro cambia poco anche assumendo altri parametri per la retta di carico, rimanendo sempre prossima (di poco superiore) a V_{thr} (in questo caso essa vale circa $V_q = 0.75$ V). Questo comportamento si può sintetizzare affermando che, approssimativamente, la d.d.p. (o caduta di potenziale) ai capi di un diodo polarizzato direttamente si mantiene sempre simile, o poco superiore, alla tensione di soglia V_{thr} . In altre parole, un diodo polarizzato direttamente provoca una caduta di potenziale che è *sempre* prossima a V_{thr} (o di poco superiore).

In genere la polarizzazione diretta di un diodo viene realizzata con un circuito in cui viene posta una resistenza R in serie tra generatore e diodo stesso. Evidentemente in questo caso la retta di carico, costruita come prima, ha intercetta con l'asse delle ordinate nel punto $V_0/(R+r)$ e quindi la pendenza della retta di carico può essere aggiustata agendo sul valore di R : in termini assoluti, essa diminuisce all'aumentare di R (supponendo V_0

e r costanti), consentendo di spostare il punto di lavoro I_q , V_q secondo le necessità.

B. Resistenza dinamica della giunzione p-n

In un componente ohmico la resistenza è definita a prescindere dalla condizioni di lavoro del componente stesso. In particolare, almeno per i regimi di frequenze di interesse per le nostre esperienze, la resistenza dei tipici resistori (non a filo) non si modifica apprezzabilmente se invece di lavorare in corrente continua si opera in alternata.

Nel caso della giunzione p-n, invece, è possibile individuare una resistenza effettiva nel caso di corrente variabile nel tempo che è *diversa* rispetto a quella si ha in corrente continua. Questo aspetto è tutt'altro che irrilevante in elettronica, dato che molto spesso una giunzione è interessata da piccole variazioni alternate di d.d.p., o corrente, e la risposta a queste variazioni alternate è importante per analizzare il comportamento dei circuiti in cui la giunzione è inserita (lo verificheremo praticamente nell'esperienza con il transistor bipolare a giunzione).

Vediamo di capire qualitativamente il perché facendo riferimento ancora all'Eq. 15, rappresentata per esempio in Fig. 9. Immaginiamo che la giunzione sia polarizzata direttamente proprio come indicato in figura, cioè che in essa scorra una certa corrente continua I_q in corrispondenza di una certa d.d.p. continua, che chiameremo qui V_q (per curiosità, il pedice q sta per *quiescent*, cioè a riposo, ovvero senza perturbazioni dipendenti dal tempo).

Supponiamo ora che a V_q sia sovrapposta una perturbazione dipendente dal tempo, che, per semplicità e anche in accordo con i metodi sperimentali di misura, sceglieremo nella forma di una *piccola* d.d.p. alternata e periodica (non interessa soffermarsi sulla frequenza e neanche sulla forma d'onda). Coerentemente con una simbologia in uso comune in elettronica, indicheremo le grandezze caratteristiche dei *piccoli* valori alternati (*ampiezze* della d.d.p. e dell'intensità di corrente associata) con le *lettere minuscole* v e i . Dunque ai capi del diodo supporremo di avere una d.d.p. variabile periodicamente nel tempo compresa fra $V - v/2$ e $V + v/2$ con $V = V_q$ costante, e la corrente che fluisce nel diodo avrà intensità oscillante tra $I - i/2$ e $I + i/2$, con $I = I_q$ costante.

A causa della ripidità della curva caratteristica del diodo, la piccola (variazione di) ampiezza di d.d.p. v potrà produrre una relativamente grande variazione di ampiezza di corrente i . Di conseguenza, il rapporto v/i potrà essere ben diverso dal rapporto V_q/I_q , determinato in corrente continua.

Definiamo *resistenza dinamica* r_d della giunzione, ovvero del diodo che ne fa uso, il rapporto $r_d = v/i$. Linearizzando al primo ordine la curva di risposta attorno al punto di lavoro V_q , I_q questo rapporto equivale al primo ordine al reciproco della pendenza della retta *tangente alla curva caratteristica* tracciata per il punto di lavoro

considerato:

$$r_d \simeq \left(\left| \frac{\partial I}{\partial V} \right|_{I=I_q} \right)^{-1}, \quad (21)$$

dove abbiamo indicato con ∂V la variazione infinitesima della d.d.p. ΔV ai capi del diodo, che al primo ordine equivale a v (e analogamente ∂I equivale a i). Viene lasciata come esercizio la dimostrazione di quanto sopra: osservate che lo sviluppo in serie di Taylor della curva di risposta permette anche di valutare quantitativamente la bontà dell'approssimazione espressa nell'Eq. 21 in funzione dell'ampiezza v della perturbazione e delle condizioni di lavoro del diodo, tutti aspetti interessanti per lo svolgimento di misure sperimentali della resistenza dinamica.

Individuare la resistenza dinamica con il reciproco della retta tangente alla curva di risposta consente di valutare immediatamente, anche se in modo approssimativo (con la semplificazione della matematica) la resistenza dinamica stessa. È evidente che se il punto di lavoro della giunzione è prossimo alla soglia, allora la pendenza (coefficiente angolare) della retta tangente può essere relativamente piccola, per cui la resistenza dinamica può essere relativamente grande. Invece se il punto di lavoro è oltre soglia, dove la curva caratteristica diventa considerevolmente ripida, la tangente può diventare grande e la resistenza dinamica piccola.

Dal punto di vista geometrico la resistenza dinamica così valutata equivale alla *cotangente* della curva, che è generalmente diversa dalla *cosecante* corrispondente alla resistenza effettiva (in continua) del diodo.

Supponendo valido il modello di Shockley, possiamo

stimare il valore della resistenza dinamica r_d . Si ha infatti

$$r_d \simeq \frac{\partial I}{\partial V} = \frac{\partial}{\partial V} I_0 \left[\exp \left(\frac{\Delta V}{\eta V_T} \right) - 1 \right] \sim \quad (22)$$

$$\sim \frac{\partial}{\partial V} I_0 \exp \left(\frac{\Delta V}{\eta V_T} \right), \quad (23)$$

dove abbiamo supposto $V > V_{thr}$ e dunque abbiamo trascurato il termine -1 rispetto all'esponenziale. Calcolando la derivata e facendo il reciproco, in accordo con la definizione, si ha

$$r_d \simeq \frac{\eta V_T}{I_0 \exp(V/(\eta V_T))} \sim \frac{\eta V_T}{I_q}, \quad (24)$$

dove abbiamo posto $I_0 \exp(\Delta V/(\eta V_T)) \sim I_q$.

La resistenza dinamica risulta correttamente dipendente dal punto di lavoro, in particolare dall'intensità di corrente di lavoro I_q : per fare un esempio, nelle condizioni di Fig. 9, dove si era supposto $I_q \simeq 12$ mA, si ha $r_d \simeq 4$ ohm; se però si supponesse una corrente di lavoro, ottenuta polarizzando in modo diverso la giunzione, $I_q \simeq 1$ mA, si avrebbe $r_d \simeq 50$ ohm, mentre per $I_q \simeq 50$ mA (valore a cui i diodi ordinari cominciano a soffrire parecchio a causa della "dissipazione Joule" sulle resistenze dei pezzi di semiconduttore, elettrodi e fili, vedi anche Fig. 8, e il conseguente riscaldamento), si avrebbe $r_d \simeq 1$ ohm, valore molto difficile da ottenere sperimentalmente proprio a causa della presenza delle resistenze in serie.

La misura della resistenza dinamica del diodo può essere condotta in modo indiretto secondo quanto applicheremo in una specifica esperienza pratica in laboratorio.

Resistenza dinamica del diodo

francesco.fuso@unipi.it

(Dated: version 5.b - FF, 6 febbraio 2020)

Questa nota illustra i contenuti dell'esercitazione pratica di rientro dalle vacanze invernali e ha lo scopo principale di rinfrescare un po' le idee alla base di definizione e misura della resistenza dinamica anche a costo di apparire ripetitiva rispetto ad altre note.

I. POLARIZZAZIONE DEL DIODO E RESISTENZA

La resistenza elettrica di un componente, o dispositivo, a due fili è definita come $R = \Delta V / I$, con ΔV differenza di potenziale applicata al componente e I intensità della corrente che lo percorre. È evidente che questa definizione, se applicata a un diodo, porta come conseguenza la necessità di specificare che il valore della resistenza così stabilita dipende dalle condizioni di funzionamento, ovvero dal suo *punto di lavoro*: per $\Delta V < V_{thr}$ la corrente che fluisce nel diodo è sempre "molto bassa" e dunque la resistenza "molto alta", mentre per $\Delta V \geq V_{thr}$, cioè in condizioni di polarizzazione diretta, la corrente può divenire "molto alta", e dunque la resistenza "molto bassa". Infatti l'intensità di corrente che fluisce in un diodo a giunzione bipolare sottoposto alla d.d.p. ΔV è regolata dalla legge di Shockley che segue un andamento fortemente non lineare.

Determinare il punto di lavoro di un diodo significa stabilire i valori di intensità di corrente e differenza di potenziale che lo interessano (rispettivamente I_q e V_q , secondo le denominazioni standard in questo contesto). A causa del legame non lineare tra tali grandezze, il problema di stabilire il punto di lavoro può non essere di immediata soluzione dal punto di vista concettuale e neanche da quello pratico. Il modo più semplice consiste nell'inserire il diodo in una maglia con un generatore di d.d.p. V_0 e una resistenza R_P in serie (la resistenza R_P comprende anche l'eventuale resistenza interna del generatore, qualora questo non fosse ideale). L'equazione della maglia è

$$V_0 = R_P I + \Delta V, \quad (1)$$

che, nel piano ΔV , I (la d.d.p. applicata al diodo è sull'asse orizzontale, l'intensità di corrente che lo attraversa su quello verticale) è rappresentata da una retta (*retta di carico*) con coefficiente angolare negativo pari a $-1/R_P$ e intercette V_0 e V_0/R_P sugli assi rispettivamente orizzontale e verticale. Il punto di lavoro è allora quello che si ottiene come intersezione di questa retta con la curva I-V del diodo, come rappresentato in Fig. 1. Il metodo descritto conduce di fatto alla "soluzione grafica" del problema di determinare intensità di corrente e d.d.p. ai capi del diodo. Naturalmente la soluzione potrebbe anche essere determinata per altra via facendo uso dell'espressione esplicita dell'equazione di Shockley. Tuttavia normalmente i parametri che compaiono in questa equazione non

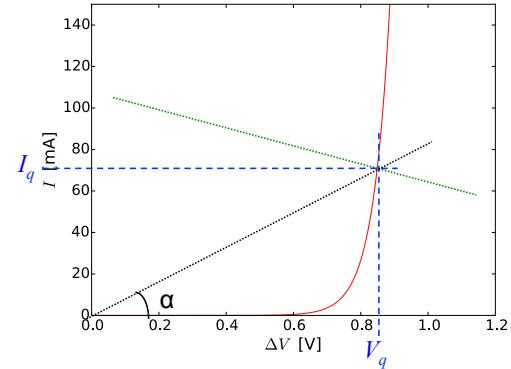


Figura 1. Curva caratteristica I-V di un ipotetico diodo con sovrapposta una altrettanto ipotetica retta di carico, che individua il punto di lavoro I_q, V_q ; la resistenza effettiva del diodo può essere definita dal rapporto V_q/I_q , che equivale al reciproco della tangente dell'angolo α indicato. Per questo esempio sono stati usati valori arbitrari di tutti i parametri.

sono noti (voi li avete determinati con un best-fit per lo specifico componente a vostra disposizione), né possono essere determinati direttamente dal datasheet del componente. Invece, almeno dal punto di vista concettuale, è sempre possibile determinare sperimentalmente la curva I vs ΔV , motivo per cui la soluzione grafica assume particolare rilevanza.

Avendo stabilito il punto di lavoro possiamo stabilire il valore della resistenza del diodo, che in questo contesto chiamiamo *resistenza "effettiva", o "efficace"*, come rapporto V_q/I_q . In termini geometrici questo rapporto indica il coefficiente angolare della retta cosecante alla curva caratteristica I-V passante per il punto di lavoro, ovvero il reciproco della tangente dell'angolo α indicato in figura.

A. Resistenza dinamica

In termini generali, la resistenza dinamica è quella che un componente oppone a una differenza di potenziale *variabile nel tempo* che ad esso è applicata. Per semplicità pratica, e senza perdere significativamente in generalità, supporremo che la d.d.p. abbia una forma sinusoidale e un'ampiezza (o ampiezza picco-picco) che, usando una convenzione frequente in elettronica, indicheremo con la

lettera minuscola v_d , dove il pedice d ci ricorda che stiamo trattando una grandezza *dinamica* (il segnale cambia nel tempo) e anche che la stiamo considerando per un diodo. Questa d.d.p. darà luogo a una corrente di intensità anche variabile nel tempo, con ampiezza (o ampiezza picco-picco) indicata dal simbolo i_d . La resistenza dinamica non è altro che il rapporto $r_d = v_d/i_d$.

Inoltre, per motivi che saranno chiari più avanti, la definizione di resistenza dinamica presuppone che la perturbazione della d.d.p. abbia un'ampiezza (o ampiezza picco-picco) v_d *piccola* (vedremo nel seguito rispetto a cosa), da cui l'uso di simboli con caratteri minuscoli. Dunque la resistenza dinamica di un diodo è definita come rapporto tra la piccola variazione di d.d.p. variabile nel tempo applicata al componente e la presumibilmente piccola variazione di intensità di corrente corrispondente. Vale la pena anticipare che l'impiego di "piccoli" segnali variabili nel tempo è tipica nei processi di amplificazione resi possibili da dispositivi "attivi" quali i transistor, all'interno dei quali è possibile trovare delle giunzioni bipolari (almeno nel transistor BJT che studieremo quest'anno).

Vediamo ora perché è lecito aspettarsi che la resistenza dinamica sia diversa da quella effettiva, o efficace, definita in precedenza. Allo scopo supponiamo, come rappresentato in Fig. 2, di avere un diodo a giunzione polarizzato in modo da operare al punto di lavoro V_q , I_q e di sovrapporre alla tensione di lavoro la "debole" d.d.p. oscillante $v_d(t)$ di "piccola" ampiezza v_d , per esempio sinusoidale e alternata. Questa darà luogo a una variazione nel tempo della corrente che attraversa la giunzione la cui intensità oscillatoria attorno a I_q . L'ampiezza di questa oscillazione è i_d . Si intuisce immediatamente da questa costruzione grafica che la resistenza dinamica $r_d = v_d/i_d$ sarà funzione del punto di lavoro e anche che essa sarà probabilmente diversa dal rapporto V_q/I_q che definisce la resistenza effettiva, o efficace.

Se l'ampiezza v_d fosse paragonabile a V_q , la resistenza dinamica sarebbe mal definita e addirittura il diodo potrebbe passare nel tempo in diverse condizioni di polarizzazione. Questo costituisce un primo banalissimo motivo per cui l'ampiezza (o ampiezza picco-picco) v_d deve essere sicuramente più piccola almeno di V_q . Inoltre possiamo anche subito notare come, supponendo sinusoidale la perturbazione sulla d.d.p., il segnale in corrente risultante non sia sinusoidale: nell'esempio mostrato in figura è evidente come l'intensità di corrente vari in modo "asimmetrico" (non alternato, cioè a media temporale non nulla).

B. Modello matematico approssimato

Immaginiamo ora di essere nel mondo, libero e semplice, della matematica. In questo contesto potremmo interpretare le ampiezze v_d e i_d come degli infinitesimi, cioè porre $v_d = dV$ e $i_d = dI$. In questo mondo la resistenza

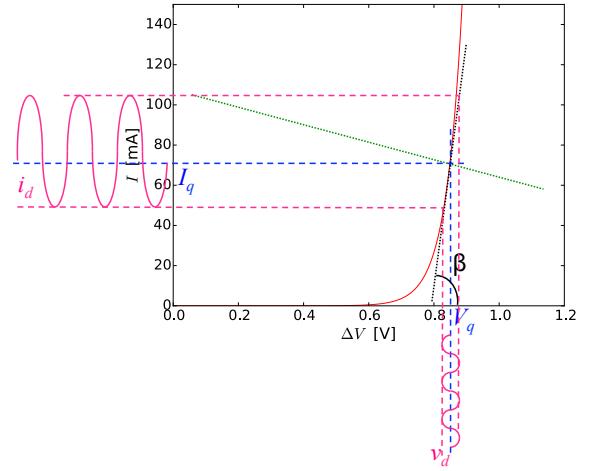


Figura 2. Illustrazione sulla determinazione grafica della resistenza dinamica di un diodo a giunzione. Il grafico mostra la stessa curva caratteristica e retta di carico di Fig. 1 e due segnali dipendenti dal tempo, corrispondenti rispettivamente alla perturbazione sulla d.d.p. e alla risultante perturbazione dell'intensità di corrente (per esigenze tipografiche le perturbazioni sono tutt'altro che piccole!). Inoltre la figura mostra la tangente alla curva di Shockley nel punto di lavoro: il suo reciproco approssima, secondo quanto discusso nel testo, la resistenza dinamica r_d .

dinamica risulterebbe

$$r_d \simeq \left. \frac{\partial V}{\partial I} \right|_{V=V_q} = \left(\frac{\partial I}{\partial V} \right)^{-1}_{I=I_q}, \quad (2)$$

dove abbiamo usato i simboli delle derivate parziali per generalità e l'ultimo passaggio è funzionale per interpretare geometricamente la resistenza dinamica come *il reciproco della tangente della curva I-V al punto di lavoro* (il reciproco della tangente dell'angolo indicato con β in Fig. 2), ovvero la cotangente al punto di lavoro.

L'Eq. 2 può essere impiegata per determinare un valore atteso per r_d (che scopriremo poi essere un valore atteso *approssimato*). A questo scopo possiamo usare l'equazione di Shockley per esprimere l'intensità di corrente I che scorre nel diodo sottoposto alla d.d.p. ΔV :

$$I = I_0 \left[\exp \left(\frac{\Delta V}{\eta V_T} \right) - 1 \right], \quad (3)$$

dove, per un diodo al silicio a temperatura ambiente, $\eta \simeq 2$ e $V_T \simeq 26$ mV, mentre I_0 rappresenta la corrente di saturazione inversa (valore tipico di alcuni nA). Nelle condizioni di interesse pratico, in cui la resistenza dinamica è valutata per giunzioni polarizzate direttamente, quindi con $\Delta V = V_q \gtrsim V_{thr} \gg V_T$, il termine esponenziale domina e l'Eq. 3 può essere bene approssimata come

$$I \simeq I_0 \exp \left(\frac{\Delta V}{\eta V_T} \right), \quad (4)$$

che, tenendo conto delle condizioni di lavoro della giunzione, porta alla seguente relazione tra I_q e V_q :

$$I_q \simeq I_0 \exp\left(\frac{V_q}{\eta V_T}\right). \quad (5)$$

Secondo l'Eq. 2 si ha

$$r_d \simeq \left(\frac{\partial I}{\partial V}\right)^{-1} = \left[\frac{I_0}{\eta V_T} \exp\left(\frac{V}{\eta V_T}\right)\right]^{-1}, \quad (6)$$

che, tenendo conto della Eq. 4 e delle condizioni di lavoro di Eq. 5, porta alla seguente espressione per il valore atteso $r_{d,att}$ della resistenza dinamica:

$$r_{d,att} \simeq \frac{\eta V_T}{I_q}. \quad (7)$$

Dunque la resistenza dinamica di una giunzione (polarizzata direttamente) è attesa essere inversamente proporzionale alla corrente di lavoro I_q attraverso un coefficiente che, per un ordinario diodo al silicio a temperatura ambiente, vale circa 52 mV/A: per esempio, per $I_q = 1$ mA, si ha $r_d \simeq 52$ ohm, che si riducono di un fattore 10 quando la corrente di lavoro diventa $I_q = 10$ mA.

Verifichiamo in che senso il valore atteso di Eq. 7 è approssimativo. In primo luogo abbiamo compiuto un'approssimazione in Eq. 4: l'errore relativo di modello che questa approssimazione comporta scala grossolanamente con V_q come $\exp[-V_q/(\eta V_T)]$ e quindi facilmente scende al di sotto delle sensibilità tipiche delle nostre misure nel caso in cui il diodo sia polarizzato direttamente.

Più rilevante, e anche più significativa dal punto di vista concettuale, è l'approssimazione inevitabilmente connessa all'uso di Eq. 2. Esprimendo r_d come la cotangente alla curva di Shockley stiamo in sostanza considerando un componente ohmico (si comporta linearmente) che ha quella determinata resistenza. In altre parole, stiamo *linearizzando* la curva stessa attorno al punto di lavoro, cioè la stiamo sviluppando al primo ordine. Si lascia per esercizio determinare l'espressione dello sviluppo e la valutazione dell'errore di modello dovuto a questa approssimazione: è facile tuttavia indovinare che il termine al secondo ordine dello sviluppo di Taylor scala come $(V_q/(\eta V_T))^2$ e che quindi l'errore relativo compiuto con il modello va come $V_q/(\eta V_T)$.

Da questa considerazione può essere anche tratta una valutazione su quanto piccola deve essere l'ampiezza (o ampiezza picco-picco) v_d . Infatti l'intero concetto di resistenza dinamica che stiamo applicando fa riferimento alla linearizzazione di cui sopra, cioè alla sostituzione del comportamento effettivo del diodo con quello di un componente ohmico. Di conseguenza è ovvio che la resistenza dinamica è ben definita solo se $v_d \ll \eta V_T$ (nell'esercitazione pratica sarà raccomandato di usare $v_d < 5$ mV_{pp}).

Infine c'è un ultimo aspetto di approssimazione di modello che riguarda la validità stessa dell'equazione di Shockley. Essa descrive, nell'ambito del modello a cui si

fa riferimento, il comportamento della sola giunzione p-n. Il trasporto di carica all'interno del diodo, però, coinvolge anche il passaggio dei portatori di carica (maggioritari, visto che si opera in polarizzazione diretta) all'interno degli spessori dei pezzi di materiale semiconduttore drogato p e n che, ovviamente, presentano mobilità finite, e attraverso elettrodi e fili di connessione, che, ovviamente, presentano resistenze non nulle. Qualitativamente è come se in serie alla giunzione si trovassero due resistenze ohmiche, una nella regione dell'anodo e l'altra in quella del catodo. Di norma, queste resistenze hanno valore (dell'ordine di qualche ohm, o anche meno) più piccolo della resistenza effettiva, o efficace, del diodo e quindi possono essere trascurate a meno di non spostare il punto di lavoro nelle zone a corrente molto alta, una situazione che nella pratica dei nostri esperimenti ordinari non si verifica. Tuttavia la resistenza dinamica può assumere valori paragonabili a qualche ohm, per cui l'effetto di queste resistenze in serie può diventare apprezzabile sperimentalmente. Inoltre la validità dei valori numerici previsti dall'equazione di Shockley dipende dalla conoscenza accurata dei parametri che essa contiene, in particolare il prodotto ηV_T , che nell'esercitazione pratica può essere preso solo come stima.

II. ESERCITAZIONE PRATICA SULLA RESISTENZA DINAMICA

Per determinare il valore di r_d è necessario conoscere le ampiezze v_d e i_d . Nell'esercitazione pratica v_d viene valutato direttamente con l'oscilloscopio, mentre la misura di i_d viene eseguita in maniera "indiretta", cioè dedotta dalla misura di una caduta di potenziale e dalla conoscenza della resistenza attraverso cui questa caduta di potenziale si verifica. Il circuito che deve essere montato è mostrato in Fig. 3(a): si vede come il diodo sia a comune tra due maglie, che chiameremo maglia di sinistra e maglia di destra.

La maglia di sinistra serve evidentemente a fornire una polarizzazione alla giunzione del diodo. Il suo punto di lavoro dipenderà dalla scelta della resistenza R_P e la corrente di lavoro I_q potrà essere letta dall'amperometro posto in serie. La maglia di destra, invece, serve per applicare al diodo la perturbazione di ampiezza v_d necessaria per la misura della resistenza dinamica e a consentirne la misura indiretta. Notiamo alcuni aspetti interessanti del circuito:

- il condensatore C serve per *disaccoppiare* le due maglie rispetto alle componenti continue. Infatti esso impedisce che la corrente continua che circola nella maglia di sinistra, necessaria per la polarizzazione del diodo, vada a finire nella maglia di destra. Se così non fosse, la lettura della corrente di lavoro eseguita dall'amperometro non rifletterebbe le effettive condizioni di operazione del diodo, dato che parte dell'intensità di corrente letta dallo strumento diramerebbe altrove.

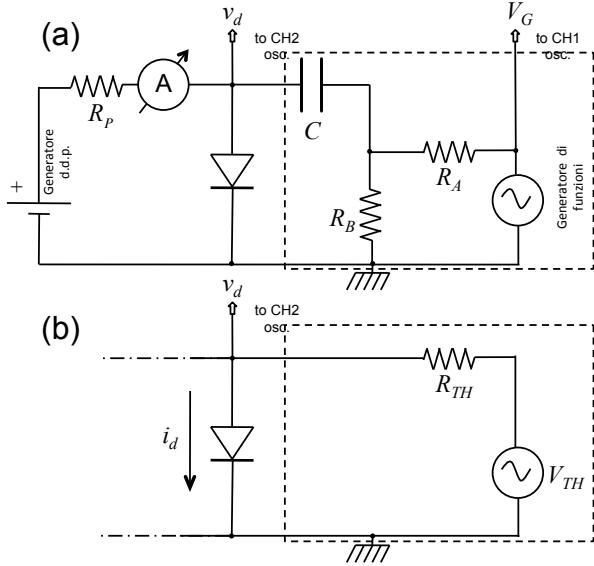


Figura 3. Schema del circuito per la misura indiretta della resistenza dinamica (a) e schema del generatore equivalente di Thévenin (b) che costituisce, nelle approssimazioni considerate nel testo, la sorgente di d.d.p. oscillante necessaria per la misura.

- La componente alternata del segnale prodotta dal generatore può invece “passare” attraverso il condensatore e dunque interessare il diodo. Se si usano i valori di capacità e frequenza di lavoro consigliati nel testo dell'esercitazione (rispettivamente $C = 10 \mu\text{F}$ nominali e $\omega = 2\pi f \sim 10^4 \text{ rad/s}$), l'impedenza del condensatore ha modulo di una decina di ohm, o minore.
- Le due resistenze R_A e R_B , che sono preassemblate in un modulo, costituiscono un partitore di tensione il cui scopo principale è quello di attenuare l'ampiezza dell'onda sinusoidale prodotta dal generatore in modo che il segnale oscillante applicato al diodo abbia un'ampiezza v_d sufficientemente piccola (deve essere $v_d \lesssim 5 \text{ mV}_{pp}$ - si ricorda anche che il generatore di funzioni in uso in laboratorio dispone al suo interno di due attenuatori da -20 dB, eventualmente collegati in cascata, azionabili agendo sulla manopola dell'ampiezza o su un apposito tastino). La scelta dei resistori del partitore ($R_A = 10R_B$, nominali) risulta in un rapporto di attenuazione nominale $R_B/(R_A + R_B) \sim 0.09$. Notate che, viste le deboli ampiezze in gioco, il rapporto S/N per la misura di v_d tramite oscilloscopio può risultare poco soddisfacente e che per migliorarlo è possibile impiegare il filtro passa-basso montato su tee-BNC, disponibile sul banco di laboratorio.
- Per la presenza del resto del circuito, in particolare del diodo che, di fatto, si trova in parallelo a R_B , il rapporto di attenuazione può raggiungere valori sensibilmente inferiori rispetto a quanto appena

scritto, ovviamente dipendenti dalle condizioni di lavoro del diodo, cioè proprio dal valore della resistenza dinamica che intendiamo misurare. In altre parole, il rapporto tra v_d e V_G (vedi figura per la simbologia) potrà diventare molto minore del rapporto di attenuazione previsto per il partitore al diminuire della resistenza dinamica del diodo. Nell'esercitazione si chiede di determinare r_d per vari valori della resistenza R , dunque per diverse condizioni di lavoro. È altamente probabile che, pur mantenendo inalterata l'ampiezza in uscita dal generatore di funzioni, indicata con V_G in Fig. 3(a) e letta dall'oscilloscopio, l'ampiezza v_d cambi. È essenziale che essa rimanga sempre nel limite indicato ($v_d \lesssim 5 \text{ mV}_{pp}$) affinché la misura abbia senso.

Dal punto di vista circuitale l'intera maglia di destra [la parte di circuito racchiusa nel box tratteggiato in Fig. 3(a)] può essere vista come un generatore *reale* di d.d.p. (alternata). Applicando l'approccio di Thévenin, possiamo dedurre la resistenza R_{TH} e la d.d.p. V_{TH} di questo generatore e considerare, per i nostri scopi, il circuito equivalente mostrato in Fig. 3(b). Per farlo, possiamo in prima battuta applicare le seguenti approssimazioni, sulla cui validità torneremo in seguito:

1. trascurare la resistenza interna $r_G = 50 \text{ ohm}$ del generatore;
2. trascurare l'impedenza del condensatore rispetto alla perturbazione variabile nel tempo;
3. trascurare completamente la corrente che fluisce nella maglia di sinistra al di fuori del diodo (cioè la corrente che eventualmente passa per la serie di amperometro, R_P e generatore di d.d.p. V_0);
4. trascurare l'effetto della resistenza interna dell'oscilloscopio in entrambi i canali.

Usando queste approssimazioni, l'approccio di Thévenin consente di valutare R_{TH} dallo schema supponendo di rimpiazzare il generatore di funzioni (ideale) con un cortocircuito. Inoltre esso permette di stabilire V_{TH} applicando le regole dei partitori di tensione alla serie di R_A e R_B alimentata dalla tensione oscillante di ampiezza V_G . La differenza tra V_{TH} e v_d (quest'ultima misurata con l'oscilloscopio) e la conoscenza di R_{TH} permettono di risalire al valore della corrente i_d , e da qui, usando la definizione, dedurre $r_d = v_d/i_d$.

In definitiva, misurando le ampiezze (o ampiezze picco-picco, è lo stesso ma si deve essere coerenti nella scelta) v_d e V_G , conoscendo attraverso misura con il tester le resistenze R_A e R_B , e supponendo valide le approssimazioni elencate prima, si ottiene una relazione complicata che permette di stabilire r_d a partire dalle grandezze misurate. Riporto per vostra referenza la relazione che esce a me:

$$r_d = \frac{v_d R_A R_B}{V_G R_B - v_d (R_A + R_B)} . \quad (8)$$

A. Incertezze e propagazione

Nella scheda che accompagna l'esperienza si chiede, per puri scopi “didattici”, di determinare passo dopo passo tutte le grandezze impiegate nel calcolo di r_d . Dovrete quindi esplicitare l'espressione di R_{TH} in funzione di R_A e R_B , quella di V_{TH} in funzione di R_A , R_B e V_G , quella di i_d in funzione di R_{TH} e V_{TH} , e infine scrivere r_d in funzione di v_d e i_d . Poiché nelle espressioni corrispondenti compaiono, come è ovvio, delle grandezze misurate, è opportuno determinare passo dopo passo i *valori* delle grandezze corrispondenti, i quali devono altrettanto ovviamente essere accompagnati dalle loro incertezze, determinate attraverso opportuna propagazione delle incertezze di misura.

In altre parole, nella scheda sarete chiamati a valutare R_{TH} con il corrispondente ΔR_{TH} , V_{TH} con il corrispondente ΔV_{TH} , e infine i_d con il corrispondente Δi_d . Potrete dunque essere “invogliati” ad esprimere l'incertezza sulla misura di r_d , Δr_d , combinando in modo opportuno le varie incertezze sulle grandezze determinate passo dopo passo. Questa procedura è ragionevole e piuttosto immediata. Notate, però, che molto probabilmente essa condurrà a una *sovraffima* dell'incertezza su r_d . Infatti alcuni termini, accompagnati dalle proprie incertezze, compaiono più volte all'interno delle varie espressioni.

Nel mio caso, in presenza di una corrente $I_q \simeq 1.3$ mA, a cui corrisponde una resistenza dinamica attesa $r_{d,att} \simeq 40$ ohm, ho ottenuto $r_d = (44.6 \pm 8.5)$ ohm. Invece, applicando la propagazione dell'errore direttamente all'espressione di Eq. 8, ho ottenuto $r_d = (44.6 \pm 5.8)$ ohm, con un'incertezza relativa che è quasi la metà di quella ottenuta prima. Dunque nell'esercitazione pratica dovete determinare l'incertezza Δr_d usando sia la propagazione dell'errore (massimo) a partire da ΔR_{TH} , ΔV_{TH} e Δv_d , sia la propagazione dell'errore (massimo) sull'espressione che lega r_d alle grandezze direttamente misurate, e che quindi utilizza “in un sol colpo” ΔR_A , ΔR_B , ΔV_G e Δv_d , almeno per una scelta di R_P .

B. Approssimazioni circuituali

L'esercitazione pratica richiede di stimare la validità delle approssimazioni impiegate per “risolvere” il circuito, in particolare di quelle listate in Sezione II. Qui, senza entrare nei dettagli che vengono lasciati per esercizio o per ulteriore discussione, presentiamo alcuni aspetti rilevanti in questo contesto.

1. In termini generali, trascurare la resistenza interna r_G del generatore di funzioni è tanto meglio giustificato quanto minore è l'intensità di corrente i_d richiesta al generatore stesso. Le condizioni “peggiori” si hanno quando r_d è piccola, cioè quando la corrente di lavoro I_q è grande, ovvero quando la

resistenza R_P è piccola. Ho fatto una stima, che invito anche voi a fare, e riscontrato che anche nelle condizioni “peggiori” la caduta di potenziale su r_G è almeno un ordine di grandezza inferiore rispetto a quella su R_{TH} . Trascurare la resistenza interna del generatore ha effetto nella determinazione di R_{TH} secondo l'approccio di Thévenin, ma non incide direttamente sulla valutazione di i_d , dato che la grandezza misurata V_G è acquisita a valle della resistenza interna, cioè all'uscita del generatore *reale*. Un modo alternativo per verificare quanto r_G sia trascurabile può essere il confronto con R_{TH} (stimato trascurando r_G), che dimostra come r_G sia ancora oltre un ordine di grandezza inferiore a R_{TH} , originando conseguenze ragionevolmente trascurabili.

2. L'impedenza del condensatore si trova in serie al diodo, per cui essa entra direttamente nella relazione che lega V_{TH} a v_d . Trascurando il condensatore, si ha $i_d = (V_{TH} - v_d)/R_{TH}$. Se il condensatore non viene trascurato, tenendo conto che in questa trattazione siamo interessati alle ampiezze dei segnali (dunque conta il modulo dell'impedenza), si ha $i_d = (V_{TH} - v_d)/(R_{TH} + |Z_C|)$. D'altra parte, $R_{TH} \simeq 620$ ohm, mentre $|Z_C| = 1/(\omega C) \simeq 11$ ohm (considerando $C = 10 \mu\text{F}$ nominali e $f \simeq 1.4$ kHz, come ho usato io), che soddisfa piuttosto bene l'approssimazione, dato che $|Z_C|/R_{TH} < 1/50$.
3. La presenza dei componenti della maglia di sinistra del circuito può essere modellata supponendo che in parallelo al diodo si trovi una resistenza effettiva R' data dalla serie della resistenza interna dell'amperometro, della resistenza R_P e della resistenza interna del generatore di d.d.p. V_0 , r_G . Stimiamo separatamente R' per i casi di alta e bassa corrente di lavoro del diodo. Ad alta corrente di lavoro, dove possiamo stimare $r_{d,att} \sim 5$ ohm, si ha per esempio $R_P = 330$ ohm (nominali). Questa resistenza è, da sola, circa 60 volte maggiore di $r_{d,att}$, per cui si può piuttosto tranquillamente assumere che la presenza di R' in parallelo al diodo sia trascurabile. I più bassi valori della corrente di lavoro consentiti nell'esperienza si ottengono ad esempio per $R = 6.8$ kohm (nominali). In queste condizioni si misurano I_q di alcune centinaia di μA , che corrispondono a valori attesi $r_{d,att}$ dell'ordine delle centinaia di ohm. Anche in questo caso è quindi ragionevole trascurare gli effetti della resistenza R' .
4. L'oscilloscopio è impiegato per misurare V_G e v_d . In entrambi i casi i suoi effetti possono essere considerati trascurabili, dato che essi possono modellati con delle resistenze molto alte (1 Mohm) poste in parallelo al generatore di funzioni e al diodo, che hanno per conto proprio delle resistenze ben minori.

Cenni sul transistor bipolare a giunzione - nuova versione

francesco.fuso@unipi.it

(Dated: version 9.b - FF, 1 marzo 2020)

Quella del transistor è probabilmente una delle scoperte più importanti che la fisica ha conseguito nel secolo scorso, almeno in termini di ricadute pratiche. Vista l'importanza dell'argomento, esistono voluminosi trattati che lo riguardano. Qui ci limitiamo a puntualizzare alcuni aspetti costruttivi e funzionali del transistor bipolare a giunzione (*BJT*) a un approccio di elettrostatica classica e a un modello microscopico qualitativo. In ogni caso, dato il carattere necessariamente stringato e semplificato di questa nota, siete invitati a riferirvi a uno dei tantissimi testi di elettronica che si trovano in biblioteca, o in rete, per chiarire eventuali punti oscuri e approfondire l'argomento.

I. INTRODUZIONE

Il transistor *BJT* rappresenta il paradigma dei dispositivi *attivi* usati in elettronica, dove l'aggettivo si riferisce alla possibilità di *controllare* qualcosa (potenziali, correnti) con qualcos'altro (altri potenziali, altre correnti) nell'ambito dello stesso componente. Allo stato attuale, a causa di limitazioni nel funzionamento e nelle tecnologie, il transistor *BJT* in quanto tale ha una diffusione piuttosto limitata rispetto ad altre tipologie di dispositivi attivi. Di certo esso si incontra meno frequentemente dei transistor di tipo *MOS-FET*, che sono presenti in milioni o miliardi di unità all'interno di qualsiasi dispositivo elettronico di tipo digitale (computer, telefonini, etc.).

Tuttavia, a parte il fatto che nell'esercitazione pratica di laboratorio si usa un transistor *BJT*, ci sono vari ottimi motivi che spingono verso l'analisi e l'interpretazione, anche semplificata, del suo funzionamento. In quanto paradigma, esso si presta infatti piuttosto bene a mostrare cosa si intende con dispositivo attivo e come un dispositivo attivo può essere impiegato praticamente.

II. DOPPIA GIUNZIONE

Il concetto di base di un transistor *BJT* (d'ora in avanti solo transistor) è la presenza di una "doppia giunzione". Le due giunzioni, ad esempio dello stesso tipo di quelle di un diodo bipolare al silicio, sono messe in serie una dopo l'altra. Dal punto di vista schematico possiamo immaginare di avere una successione di tre semiconduttori drogati in modo diverso: si può avere una successione di tipo *npn* oppure di tipo *pnp*, come rappresentato in Fig. 1, che mostra anche il simbolo circuitale per le due varianti. In questa nota faremo riferimento alla variante *npn*, sia perché di questo tipo è il transistor usato nella pratica (modello 2N1711, o equivalente), sia perché di fatto la tipologia *npn* è quella di impiego più frequente e dotata in genere di caratteristiche migliori.

La presenza delle due giunzioni potrebbe far pensare a due diodi collegati tra di loro in un montaggio "back-to-back", cioè anodo-anodo o catodo-catodo, ma questa descrizione non è sufficiente per spiegare l'effetto di "trans-

resistenza" (trans-resistor, da cui il nome di transistor) specifico del dispositivo.

Prima di esaminare queste caratteristiche specifiche soffermiamoci su alcune considerazioni molto generali:

- se abbiamo due giunzioni è evidente che possiamo ipotizzarle direttamente o inversamente l'una indipendentemente dall'altra. Nel seguito vedremo un po' meglio cosa si verifica nelle varie situazioni. Per il momento, possiamo dare un nome alle tre possibilità di interesse pratico: si dice che il transistor è *in interdizione* quando le due giunzioni sono entrambe polarizzate inversamente, *in saturazione* quando tutte e due sono polarizzate direttamente, *in regime attivo* quando una è polarizzata inversamente e l'altra direttamente (vedremo in seguito quale giunzione si trova in una condizione e quale nell'altra).
- Come precisazione di quanto appena affermato, è bene specificare che la condizione di polarizzazione diretta significa, in questo contesto, che la differenza di potenziale applicata ha la polarità "giusta" ed è sopra al valore di soglia ($V_{thr} \sim 0.45 - 0.65$ V per giunzioni bipolarie in silicio).
- Il transistor è un componente con tre elettrodi (tre filini, o reofori, che portano o prendono corrente) che hanno dei nomi caratteristici di origine "storica": *emettitore* (*E*), *base* (*B*), *collettore* (*C*), indicati anche in Fig. 1.
- Grazie ai tre elettrodi è possibile inviare al componente due segnali ("ingresso" e "uscita", o "controllore" e "controllato") riferendoli alla stessa linea. Nell'uso pratico uno dei tre elettrodi deve essere messo in comune fra "ingresso" e "uscita". Infatti è possibile montare un transistor nelle configurazioni cosiddette a *emettitore comune*, *base comune*, *collettore comune*. Qui non esamineremo l'intera casistica nei dettagli, limitandoci a una descrizione un po' più completa della sola configurazione a emettitore comune.
- Infine c'è un'ulteriore conseguenza del fatto che il transistor è un componente con tre elettrodi, dunque diverso dagli altri incontrati in precedenza. In

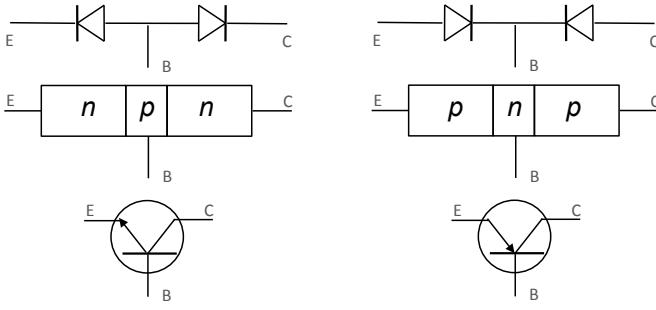


Figura 1. Rappresentazione schematica della doppia giunzione che si trova in un transistor BJT di tipo *npn* o *pnp* e simboli circuituali (le lettere E, B, C si riferiscono ai tre elettrodi del dispositivo).

quei componenti (ad esempio resistori, ma anche diodi) era possibile ipotizzare la presenza di semplici relazioni in grado di legare le due grandezze di interesse, per esempio intensità di corrente attraverso il dispositivo e differenza di potenziale applicata. Qui, invece, sarà necessario trovare delle forme più complicate per rappresentare o predire il comportamento, come ad esempio i grafici delle curve (notate il plurale) di risposta.

III. EFFETTO TRANSISTOR

Il semplice disegno rappresentato in Fig. 1 non tiene conto dell'effettiva realtà costruttiva dei transistor. Infatti:

- le tre regioni di materiale con diverso drogaggio non hanno le stesse dimensioni: la base (la regione drogata *p* in un transistor *npn*) può essere considerata molto sottile, sicuramente molto più della regione corrispondente al collettore (lo spessore in questione è in genere sub-micrometrico);
- il drogaggio delle tre regioni non è lo stesso (in termini di densità di drogante: è ovvio che nella regione *n* il drogaggio è con donori e nella *p* è con accettori): in particolare l'emettitore è molto più drogato della base e, in genere, anche del collettore (l'elevata densità di drogaggio fa spesso indicare il materiale come n^+);
- la geometria rappresentata in Fig. 1, che è sicuramente molto poco probabile dal punto di vista tecnologico (immaginate di inserire un numero gigantesco di bacheche di quel tipo all'interno di un chip...), non rappresenta affatto la realtà costruttiva. Un esempio più ragionevole è mostrato in Fig. 2(a), dove si vede uno schema compatibile con la tecnologia *planare* in uso nella microelettronica e da cui si può capire che, di fatto, l'emettitore

è circondato dal materiale della base, a sua volta circondata dal materiale del collettore.

Queste caratteristiche costruttive determinano una forte interazione tra la corrente che interessa le due giunzioni. Per capire quali siano le conseguenze di questa interazione facciamo riferimento allo schema di Fig. 2(b), dove si vedono due distinti generatori di differenza di potenziale continua collegati in modo da polarizzare *direttamente la giunzione BE* attraverso la d.d.p. V_{BE} e *inversamente la giunzione BC* attraverso la d.d.p. V_{CB} . Questo si capisce facilmente ricordando che polarizzare direttamente una giunzione *np* richiede di collegare il polo negativo del generatore alla regione *n* e quello positivo alla *p* (diamo qui per scontato che i generatori erogino delle d.d.p. superiori, in modulo, a V_{thr}), e viceversa.

In queste condizioni il transistor lavora nel *regime attivo* e c'è un flusso di elettroni, portatori maggioritari nella regione di emettitore, che il generatore V_{BE} invia alla giunzione emettitore-base. Se si trattasse di una giunzione "isolata", cioè di un semplice diodo, questo flusso di elettroni che entra (viene "emesso") dall'emettitore si ritroverebbe a uscire dalla base. Invece l'effetto transistor fa sì che una gran parte di questo flusso scavalchi la giunzione tra base e collettore e quindi venga "raccolto" dal collettore.

Questo si verifica fondamentalmente per tre motivi principali:

1. il drogaggio *p* della base ha una densità molto minore del drogaggio *n* (o n^+) dell'emettitore: il grande (denso) flusso di elettroni che si muove nell'emettitore e supera l'interfaccia tra emettitore e base non può ricombinarsi in modo "completo" nella regione di base, dove non trova una sufficiente densità, o quantità, di lacune disponibili. Pertanto non c'è un consistente richiamo di cariche positive (lacune) dall'elettrodo di base, cioè la corrente di base non ha la stessa intensità di quella di emettitore, come si verificherebbe in un diodo.
2. A causa della polarizzazione inversa, nella regione del collettore esiste un campo elettrico che accelera gli elettroni che superano la barriera BC in modo che essi possano percorrere la regione del collettore, che è drogata *n*, e di qui fuoriuscire.
3. Forma e spessore del dispositivo aiutano a rendere particolarmente efficace l'effetto: infatti lo spessore ridotto della regione di base permette agli elettroni di essere facilmente catturati dal campo presente nel collettore (ricordate anche che la polarizzazione inversa della giunzione BC implica una regione di svuotamento che si estende ben all'interno della regione di base) e la geometria del dispositivo [vedi Fig. 2(a)] rende il processo quasi isotropo spazialmente (non esistono direzioni in cui il processo sia meno efficace).

Riassumiamo il processo seguendo il flusso di elettroni che vengono iniettati nell'emettitore dove sono portatori

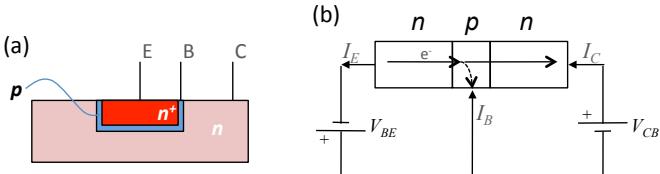


Figura 2. Geometria costruttiva tipica per un transistor *BJT* *npn* (a); rappresentazione schematica del passaggio di elettroni dall’emettitore al collettore responsabile per l’effetto transistor.

maggioritari. Inizialmente essi si muovono nella regione dell’emettitore per effetto del campo elettrico prodotto dalla d.d.p. V_{BE} . Giunti all’interfaccia con la base, in minima parte ricombinano con le lacune, richiamando una piccola quantità di cariche positive dall’elettrodo di base, cioè dando luogo a una *piccola* corrente I_B (entrante nel dispositivo) fatta di lacune (portatori maggioritari nella regione di base). La parte che non riesce a ricombinare attraversa la regione di svuotamento della giunzione BC, che ha uno spessore relativamente grande a causa della sua polarizzazione inversa ed è sede di un campo elettrico che ha verso tale da spostare gli elettroni nella direzione del collettore, transita nel collettore e fuoriesce dall’elettrodo di collettore, dando luogo alla corrente I_C (entrante nel componente).

A. Correnti nel transistor

Cerchiamo ora di quantificare l’effetto transistor. Allo scopo definiamo I_E , I_B , I_C le intensità di corrente che scorrono rispettivamente verso l’emettitore, la base e il collettore. Per evitare problemi con i segni, seguiamo la convenzione per la quale le correnti sono positive se entrano nel dispositivo [si vedano le frecce riportate in Fig. 2(b)], negative se ne escono. Notate che, secondo questa convenzione, le correnti di emettitore e collettore hanno segni opposti: infatti nelle condizioni di Fig. 2(b) si ha $I_E < 0$ (entrano elettroni nell’emettitore, corrispondenti a una corrente, fatta di cariche positive per definizione, che esce dall’emettitore stesso) e $I_C > 0$ (escono elettroni dal collettore, e quindi la corrente entra nel dispositivo).

L’effetto transistor è riassunto in questa semplice equazione:

$$I_C = -\alpha_F I_E , \quad (1)$$

dove il coefficiente α_F ha un valore prossimo all’unità: a seconda del tipo di transistor, esso è infatti tipicamente compreso tra 0.95 e 0.999. Questo significa che quasi la totalità della corrente (fatta di elettroni) che entra e si muove nell’emettitore preferisce farsi raccogliere, e quindi uscire, dal collettore piuttosto che ricombinare nella base e dare origine a una corrente di base (fatta di lacune).

Come già affermato la serie delle due giunzioni *npn* che formano il transistor *non è affatto simmetrica*: in altre

parole, emettitore e collettore non possono scambiarsi di ruolo. Infatti se si prova a “invertire” il sistema, cioè a polarizzare inversamente la giunzione BE e direttamente quella BC [si tratta, in sostanza, di scambiare i segni dei due generatori di d.d.p. che compaiono in Fig. 2(b)], si ottiene che l’effetto transistor, seppur ancora presente, è nettamente meno importante. In particolare in queste condizioni si ha $I_E = -\alpha_R I_C$, con α_R tipicamente dell’ordine di 0.5, o inferiore. Notate il pedice “R” che sta per “reverse” così come il pedice “F” usato prima stava per “forward”.

B. Amplificazione di corrente

Le tre correnti che entrano nel componente attraverso emettitore, base e collettore devono dare somma nulla. Infatti all’interno del transistor non esiste alcuna sorgente, o pozzo, di cariche. Quindi il transistor può essere considerato come un nodo al quale applicare una delle cosiddette leggi di Kirchoff. Si ha allora

$$I_E + I_B + I_C = 0 . \quad (2)$$

Utilizzando la relazione di Eq. 1 e rimaneggiando si ottiene facilmente

$$I_B = -(I_C + I_E) = \left(\frac{1}{\alpha_F} - 1 \right) I_C = \frac{1 - \alpha_F}{\alpha_F} I_C , \quad (3)$$

ovvero

$$I_C = \frac{\alpha_F}{1 - \alpha_F} I_B = \beta_F I_B . \quad (4)$$

Questa equazione dimostra due aspetti fondamentali:

1. nelle condizioni che stiamo considerando, cioè quando il transistor si trova ad operare in regime attivo (polarizzazione diretta per la giunzione BE, inversa per la BC), la corrente di collettore dipende *linearmente* dalla corrente di base e, in prima approssimazione, *solamente* da questa.
2. Poiché $\alpha_F \simeq 1$, il coefficiente β_F , talvolta indicato come h_{FE} e noto come *guadagno in corrente continua* del transistor, assume valori molto grandi, tipicamente compresi fra 50 e 1000: *la corrente di collettore è quindi amplificata rispetto a quella di base*.

In questa affermazione c’è molto del carattere “attivo” del transistor: la corrente di base controlla la corrente di collettore, cioè modificando l’una si modifica l’altra. Generalmente la corrente di base è “piccola” e quella di collettore “grande”: dunque una piccola corrente controlla una grande corrente. In condizioni opportune questo tipo di legame può essere sfruttato per ottenere un’amplificazione, cioè un guadagno in termini di correnti e/o, come vedremo, tensioni. In altre, esso può essere sfruttato per realizzare uno *switch*, cioè un interruttore in

cui il passaggio di una rilevante intensità di corrente è controllato da una ridotta intensità di corrente.

È bene rimarcare che il controllo operato dal transistor *BJT* richiede di manipolare delle correnti. Come sapete, far passare delle correnti all'interno di un dispositivo dà inevitabilmente luogo a dissipazione di potenza, poiché la resistenza non è mai nulla. La dissipazione è sempre mal vista, sia perché comporta dispendio di energia, sia perché provoca, in un modo o nell'altro, surriscaldamento. Se un chip di quelli che sono all'interno di un qualsiasi dispositivo elettronico attuale, che contengono miliardi di dispositivi di "controllo" più o meno indipendenti tra loro, fosse realizzato con la tecnologia *BJT* occorrerebbero delle ventole gigantesche e ogni ora di facebook costerebbe uno stonfo in bolletta dell'ENEL... Questo è uno dei principali motivi per cui la tecnologia *BJT* è stata ampiamente rimpiazzata, in ambito di elettronica digitale, da quella a effetto di campo (per esempio *MOS-FET*), che opera il controllo tramite applicazione di d.d.p. e non di corrente.

IV. EQUAZIONI DI EBERS-MOLL

La semplicissima Eq. 4 riporta tutto quello che è necessario sapere quando si interpreta la maggior parte delle applicazioni di un transistor. Tuttavia vale la pena di soffermarsi su un modello analitico che è in grado di chiarire un po' meglio alcuni aspetti del funzionamento. Tale modello, detto di Ebers-Moll, fa uso di equazioni che sono formalmente simili a quelle del modello di Shockley per le giunzioni bipolari in semiconduttori.

Se non ci fosse l'effetto transistor, cioè se il nostro dispositivo fosse completamente interpretabile come una coppia di diodi "back to back", allora le correnti di emettitore e collettore andrebbero scritte come

$$I_E = -I_{0E} \left[\exp\left(\frac{V_{BE}}{\eta V_T}\right) - 1 \right] \quad (5)$$

$$I_C = -I_{0C} \left[\exp\left(\frac{V_{BC}}{\eta V_T}\right) - 1 \right], \quad (6)$$

dove I_{0E} e I_{0C} sono le correnti di saturazione inversa per le due giunzioni BE e BC, $V_T = k_B T/e$ è la differenza di potenziale termica (vale circa 26 mV a temperatura ambiente) ed η è il coefficiente dovuto alle caratteristiche specifiche delle giunzioni (nel caso del diodo al silicio vale circa 2, nel caso del transistor assume generalmente un valore prossimo a 1). Riconoscete facilmente la forma delle equazioni di Shockley, qui adattate alla presenza di due distinte giunzioni. Notate anche che il segno negativo usato è dovuto alla convenzione che abbiamo dichiarato di seguire prima, per la quale le correnti sono positive se entrano nel dispositivo. Infatti tutte e due le giunzioni considerate, se polarizzate direttamente (V_{BE} e V_{BC} entrambi positivi e maggiori di V_{thr}), sostengono un flusso di elettroni entrante, che corrisponde a una corrente uscente, da cui il segno negativo.

Per l'effetto transistor queste due equazioni non sono sufficienti a descrivere il comportamento del dispositivo. Infatti è necessario aggiungere due termini ottenendo le equazioni di Ebers-Moll:

$$I_E = -I_{0E} \left[\exp\left(\frac{V_{BE}}{\eta V_T}\right) - 1 \right] + \alpha_R I_{0C} \left[\exp\left(\frac{V_{BC}}{\eta V_T}\right) - 1 \right] \quad (7)$$

$$I_C = -I_{0C} \left[\exp\left(\frac{V_{BC}}{\eta V_T}\right) - 1 \right] + \alpha_F I_{0E} \left[\exp\left(\frac{V_{BE}}{\eta V_T}\right) - 1 \right] \quad (8)$$

dove i coefficienti α_F e α_R sono quelli, già introdotti, che mescolano la corrente di collettore con quella di emettitore, e viceversa. Notate che, sulla base delle considerazioni svolte prima, i segni sono corretti.

Quando il transistor lavora in regime attivo, cioè quando la giunzione BE è polarizzata direttamente e quella BC inversamente, alcuni dei termini appena scritti diventano trascurabili. Infatti in queste condizioni $V_{BE} > V_{thr}$, quindi i termini che contengono questa tensione prevalgono sugli altri. Invece si ha $V_{BC} < V_{thr}$, per cui i termini corrispondenti possono essere trascurati, dando eventualmente luogo alla corrente di saturazione inversa.

In altre parole, in regime attivo si ha

$$I_E \simeq -I_{0E} \exp\left(\frac{V_{BE}}{\eta V_T}\right) \quad (9)$$

$$I_C \simeq \alpha_F I_{0E} \exp\left(\frac{V_{BE}}{\eta V_T}\right) = \alpha_F I_E, \quad (10)$$

che equivale sostanzialmente a quanto riportato nell'Eq. 1. Quindi, quando il transistor si trova a operare in regime attivo questo modello conduce ancora, almeno in prima approssimazione, a una corrente di collettore che è proporzionale a quella di emettitore, con un fattore di proporzionalità prossimo all'unità.

Le equazioni di Ebers-Moll, che si applicano in particolare per descrivere il regime attivo, permettono anche di capire un po' meglio cosa succede quando il transistor è in interdizione (tutte e due le giunzioni sono polarizzate inversamente) o in saturazione (tutte e due polarizzate direttamente).

In interdizione tutti i termini con l'esponenziale sono trascurabili rispetto al -1 ; quindi le correnti di emettitore e collettore corrispondono a somme di quelle di saturazione inversa, che sono generalmente così piccole da poter essere approssimate con lo zero. Allora *in interdizione non ci sono correnti* di emettitore e collettore (ovvero, nella pratica, esse sono trascurabili).

In saturazione tutti i termini con l'esponenziale prevalgono sul -1 . Dunque in questo caso *circolano correnti sia per la base che per il collettore*, e la corrente di collettore è più intensa di un fattore β_F rispetto a quella di base. Notate anche che, se $V_{BE} \approx V_{BC}$, i due termini della corrente di collettore nell'Eq. 7 tendono a elidersi, essendo praticamente uguali e di segno opposto.

Per quanto formalmente utili come modello, le equazioni di Ebers-Moll hanno un limitato impiego pratico a

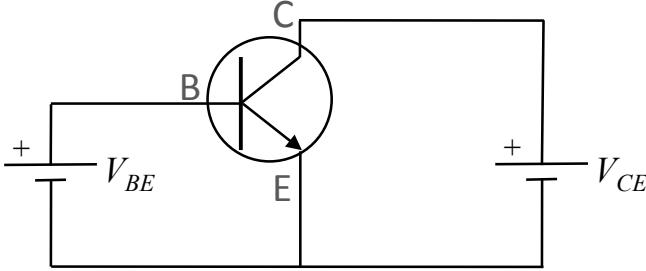


Figura 3. Schema circuitale della configurazione a emettitore comune con polarizzazione diretta della giunzione base-emettitore e inversa di quella collettore-base.

causa del gran numero di parametri che esse contengono e dalla scarsa conoscenza del valore che questi parametri possono assumere. In genere, piuttosto che usare queste equazioni, è molto più sensato impiegare le curve caratteristiche del componente specifico che si ha a disposizione.

V. CONFIGURAZIONE A EMETTITORE COMUNE E CURVE CARATTERISTICHE

Come già discusso, nell'impiego pratico di un transistor occorre riferire i potenziali a uno degli eletrodi. La Fig. 2(b) e gran parte dei discorsi fatti finora hanno avuto a che fare, almeno implicitamente, con la configurazione in cui l'elettrodo di base è quello di riferimento, cioè alla configurazione detta *a base comune*. Questo ci ha consentito di descrivere l'effetto transistor e di ottenere il guadagno in corrente β_F , che, tuttavia, sono caratteristiche del componente che permangono anche quando esso viene montato in una diversa configurazione.

Quella di maggiore interesse, sia pratico che concettuale, è probabilmente la configurazione *a emettitore comune* rappresentata schematicamente in Fig. 3. Poiché il riferimento è la linea dell'emettitore, le d.d.p. rilevanti sono qui V_{BE} e V_{CE} .

Come già anticipato, la complessità del dispositivo e del suo comportamento rendono non immediato individuare delle rappresentazioni grafiche (curve caratteristiche) che permettano di capire al primo colpo il comportamento di un transistor. Infatti le grandezze che entrano in gioco sono tante, almeno le tre tensioni V_{BE} , V_{BC} , V_{CE} e le tre correnti I_E , I_B , I_C . I legami fra queste grandezze non sono sempre ovvi.

Una delle curve caratteristiche, quella che generalmente viene chiamata *caratteristica di ingresso*, o *curva caratteristica di base*, riguarda l'andamento della corrente di base I_B in funzione della tensione V_{BE} . Questa curva possiamo facilmente indovinarla: infatti essa, riguardando sostanzialmente la sola giunzione BE, avrà l'andamento tipico alla Shockley, cioè sarà di tipo esponenziale, come rappresentato in Fig. 4(a): sotto soglia non si avrà praticamente passaggio di corrente, sopra soglia si avrà

una crescita esponenziale della corrente in funzione della tensione come in un diodo, con l'unica, ma importante, differenza che i valori (la scala) della corrente saranno qui ben minori che nel caso del diodo (μA invece di mA) a causa del fatto che il flusso di elettroni viene raccolto dal collettore invece di fluire nell'elettrodo di base. È anche chiaro che, superata la soglia, V_{BE} tenderà a rimanere costante a prescindere dal valore di I_B , o, se preferite, I_B potrà aumentare anche di parecchio senza apprezzabili aumenti di V_{BE} . Questa è una ovvia conseguenza dell'andamento esponenziale, che si verifica anche a grandi intensità di corrente a patto di trascurare, come nel diodo, la resistenza intrinseca degli elementi della giunzione.

Un altro tipo di rappresentazione che è abbastanza frequentemente usata, almeno per transistor montati a emettitore comune, è quella che riporta I_C in funzione di V_{CE} : fate bene attenzione al fatto che stavolta la differenza di potenziale non è riferita a una giunzione, ma alla serie di due giunzioni. Inoltre notate che questo tipo di rappresentazione non è univoco, nel senso che la dipendenza che si vuole rappresentare non dà luogo a una singola curva. Essa infatti è funzione anche di altri parametri "esterni" alla giunzione considerata, in particolare della corrente di base I_B . Dunque si ottiene una *famiglia di curve* in cui si riporta l'andamento di I_C in funzione di V_{CE} e ogni curva si riferisce a un dato valore di I_B [vedi Fig. 4(b)]. In genere ci si riferisce a questi grafici come a quelli delle *curve caratteristiche di uscita*, o *curve di collettore*.

Nella descrizione dell'effetto transistor data sopra, nel regime attivo I_C dipende *soltamente* (o quasi, vedi dopo) da I_B attraverso il fattore di guadagno, β_F . Allora nel regime attivo queste curve devono essere tante linee orizzontali (non c'è dipendenza da V_{CE} , cioè dalla grandezza graficata lungo l'asse orizzontale) parallele fra loro e il rapporto tra il valore di I_B corrispondente a una certa curva e quello di I_C riportato sull'asse verticale deve essere dell'ordine di grandezza di β_F . Vedremo nel seguito dove individuare i regimi di saturazione e interdizione nei grafici delle curve caratteristiche di uscita, tornando ancora sulle approssimazioni che abbiamo usato nel nostro approccio (molto) qualitativo.

A. Effetto Early (e altri dettagli)

Nella realtà, come si osserva anche in Fig. 4(b), c'è una *piccola* dipendenza residua di I_C da V_{CE} che fa sì che i tratti orizzontali siano inclinati a formare delle rette con un piccolo coefficiente angolare positivo, almeno nell'intervallo di parametri di interesse pratico. Spesso si associa questa circostanza al cosiddetto *effetto Early*. Aumentando V_{CE} si hanno diverse conseguenze all'interno del transistor, tra le quali:

- la regione di svuotamento della giunzione BC, che risulta sempre più polarizzata negativamente (in

seguito ne daremo una giustificazione quantitativa), aumenta la sua estensione longitudinale [qui si fa riferimento alla geometria lineare, quella rappresentata in Fig. 1 e Fig. 2(b)] e si ottiene un regime detto, talvolta, di *sovrasvuotamento*. Di conseguenza, gli elettroni che superano l'interfaccia emettitore-base trovano una densità minore, ovvero un numero minore, di lacune con cui ricombinarsi, diminuendo, comparativamente, la corrente di base e aumentando, comparativamente, quella di collettore.

- Inoltre il campo elettrico nella regione di collettore, che ha lo scopo di raccogliere gli elettroni che non ricombinano nella base, diventa più intenso a causa dell'aumento (in modulo) della d.d.p. che lo crea. Di conseguenza gli elettroni vengono raccolti più efficacemente dal collettore, dando luogo a una corrente di collettore più intensa.

La combinazione di questi due effetti risulta in una sorta di aumento di efficacia dell'effetto transistor, ovvero α_F tende sempre più all'unità, con un conseguente aumento di β_F . L'intensità di corrente di collettore può essere approssimata al primo ordine dalla funzione

$$I_C \sim \beta_F I_B \left(1 + \frac{V_{CE}}{V_{Early}} \right), \quad (11)$$

dove il *potenziale di Early* V_{Early} dipende dalla costruzione del transistor. Osservate che, in conseguenza dell'andamento lineare espresso dall'Eq. 11, $-V_{Early}$ può essere interpretato come il valore dell'intercetta della curva caratteristica (nel tratto considerato, approssimato a lineare) con l'asse orizzontale del grafico. Tipicamente V_{Early} vale diverse decine, o addirittura centinaia, di V, per cui la pendenza di questi tratti quasi-lineari risulta molto poco pronunciata.

L'efficacia dell'effetto transistor, oltre a dipendere da V_{CE} , può anche essere funzione di tutte le altre tensioni, o correnti, coinvolte nell'operazione del transistor, e anche dalle effettive condizioni di operazione, per esempio la temperatura alla quale si trovano le giunzioni. A complicare ulteriormente lo scenario, occorre ribadire che gli ordinari processi di fabbricazione dei transistor sono in grado di definire il valore di α_F (e anche di α_R) in modo piuttosto grossolano, cioè con una notevole tolleranza produttiva. In conclusione, possiamo sicuramente affermare che la descrizione che abbiamo dato, in cui i parametri dell'effetto transistor si considerano costanti, può risultare parecchio approssimativa. Tuttavia essa è più che adeguata per comprendere gli aspetti fondamentali del funzionamento del transistor e delle sue applicazioni, per cui continueremo a basarci largamente su di essa.

B. Retta di carico e punto di lavoro

Secondo quanto abbiamo discusso, le condizioni di lavoro del transistor dipendono dalle tensioni applicate alle

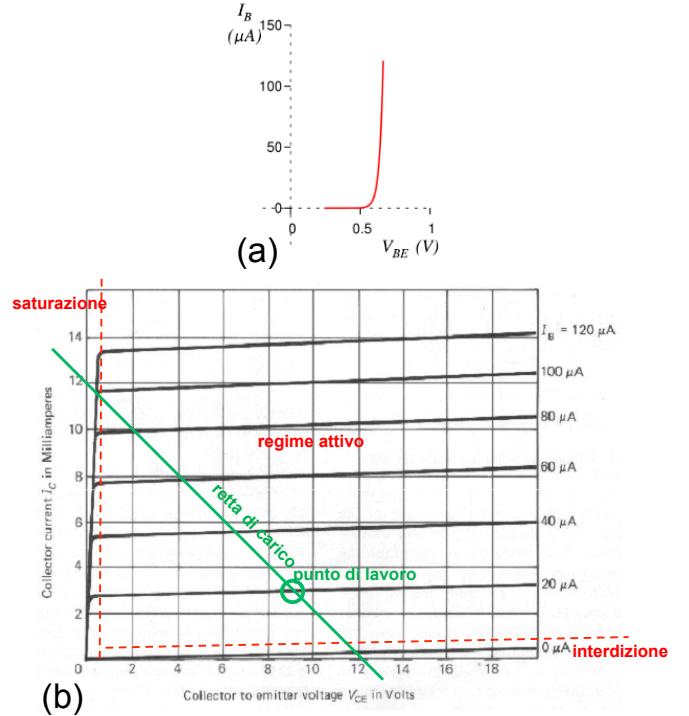


Figura 4. Curva caratteristica di ingresso (a) e di uscita (b) (dette anche curve di base e di collettore, o anche di trasferimento diretto e di uscita) per un transistor *npn*. Sul grafico delle curve di uscita sono indicate grossolanamente le zone corrispondenti ai vari regimi di funzionamento del transistor e sono sovrapposti retta di carico e punto di lavoro, individuati supponendo come esempio $V_0 = 12$ V, $R_C = 1$ kohm, $I_B = 20 \mu\text{A}$.

due giunzioni, le quali determinano anche le correnti che scorrono nel componente e, in definitiva, il *punto di lavoro* del transistor. Normalmente il regime operativo del transistor viene realizzato usando un solo generatore di differenza di potenziale continua, invece dei due generatori indipendenti che abbiamo ipotizzato finora (vedi Fig. 3, ma anche Fig. 2).

Uno schema possibile, super-semplificato (e per questo, in realtà, poco comune, discuteremo poi alcuni suoi punti deboli) è quello di Fig. 5(a): un opportuno collegamento di resistori consente di ottenere la corretta polarizzazione della giunzione BE, diretta, e della giunzione BC, inversa. La corrente I_B scorre attraverso la serie dei resistori R_P e R_B , la corrente I_C scorre attraverso il resistore R_C . Dimensionando opportunamente il valore delle resistenze si varia la d.d.p. V_{BE} , e dunque si stabilisce l'eventuale funzionamento del transistor in regime attivo. Per individuare l'effettivo punto di lavoro della giunzione BE conoscendo V_0 (qui talvolta indicato come V_{CC}) si può procedere come nel caso del diodo, cioè costruire la retta di carico corrispondente alla scelta di R_P (ovvero della serie $R_P + R_B$) e verificare dove essa intercetta la curva caratteristica di base.

Per determinare completamente il punto di lavoro del

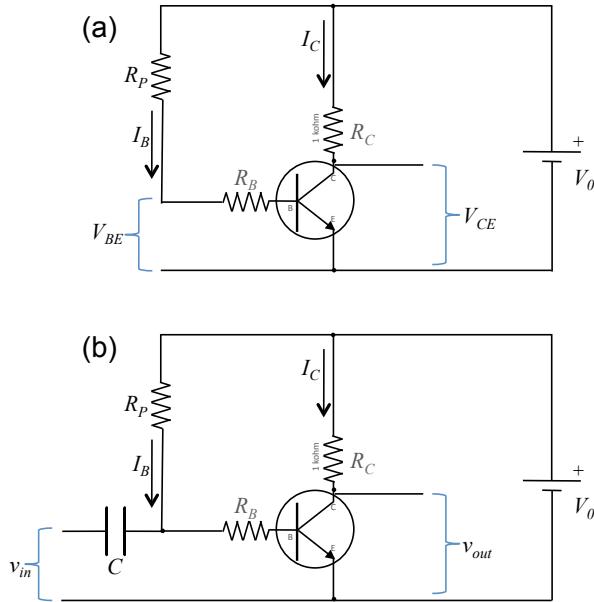


Figura 5. Esempio di semplice circuito di polarizzazione del transistor nella configurazione a emettitore comune (a); variante del circuito che prevede segnali alternati in ingresso e uscita (b) e che realizza un amplificatore di corrente e tensione, secondo quanto specificato nel testo.

transistor occorre però preoccuparsi anche di I_C , e quindi costruire una diversa retta di carico, stavolta sul piano I_C, V_{CE} . Esaminiamo la “maglia del collettore”, costituita dal generatore V_0 (supposto ideale), dalla resistenza R_C e dal transistor. Notiamo che in questa maglia passa la maggior parte della corrente che viene fornita dal generatore: infatti nella “maglia di base” scorre normalmente una corrente molto bassa (β_F -volte minore di quella che scorre nel collettore). L’equazione della maglia si può allora scrivere con buona approssimazione come

$$V_0 = V_{CE} + R_C I_C , \quad (12)$$

che, sul piano V_{CE}, I_C , dà luogo a una retta la quale vincola I_C a V_{CE} . Le intercette di questa retta con gli assi sono ovviamente V_0 e V_0/R_C , per cui la retta può essere disegnata in modo immediato conoscendo i valori rilevanti. Per esempio, se a parità di V_0 R_C viene fatta aumentare, o diminuire, l’intercetta con l’asse verticale si abbassa, o si alza, e la retta diventa meno, o più, inclinata, determinando un diverso punto di lavoro per il transistor. Il punto di lavoro del transistor sarà allora determinato dall’intersezione fra la retta di carico e la curva caratteristica sul piano I_C, V_{CE} che corrisponde al valore di I_B nelle condizioni specifiche di funzionamento del transistor.

C. Interdizione e saturazione

Vediamo ora di individuare nei grafici di Fig. 4 i regimi di interdizione e di saturazione. In interdizione la giun-

zione BE è polarizzata inversamente, cioè $V_{BE} < V_{thr}$, per cui $I_B \simeq 0$. Dunque il regime di interdizione corrisponde alle curve con $I_B \simeq 0$ e in queste condizioni, giustamente, $I_C \simeq 0$ a prescindere dal valore di V_{CE} .

A saturazione entrambe le giunzioni sono polarizzate direttamente. Essendo il transistor considerato di tipo *npn*, affinché la giunzione BC sia polarizzata direttamente occorre che il potenziale al collettore sia minore di quello alla base (entrambi i potenziali sono riferiti alla stessa linea di massa, quella a cui è collegato l’emettitore). Infatti, usando una simbologia di significato immediato, $V_{CE} = V_C - V_E = V_C - V_B + V_B - V_E = -V_{BC} + V_{BE}$. D’altra parte deve essere $V_{BE} \gtrsim V_{thr}$ se si vuole anche la giunzione BE polarizzata direttamente. Quindi per accedere al regime di saturazione V_{CE} deve essere piccola; nella pratica occorre $V_{CE} < V' < V_{thr}$, con V' un valore che dipende dallo specifico transistor utilizzato e dalle sue condizioni di funzionamento. Nel grafico di Fig. 4(b) il regime di saturazione corrisponde alla parte “di sinistra”, dove si vede che V' dipende dalla I_B , rimanendo sempre $V' < V_{thr}$.

Può essere utile a questo punto mettere in chiaro una precisazione che, in realtà, è stata data per sottointesa visto il carattere di ovietà. È evidente, credo per tutti, che le nozioni di saturazione e interdizione devono essere prese con una certa “flessibilità”. Infatti esse hanno a che fare con le condizioni di polarizzazione, diretta o inversa, delle giunzioni. Dal punto di vista modellistico è abbastanza facile definire la polarizzazione di una giunzione come diretta o inversa. Dal punto di vista pratico le situazioni sono più confuse. Di conseguenza, spesso si assume che il regime di operazione di un transistor sia in saturazione o in interdizione quando esso *si avvicina* alla saturazione o all’interdizione, cioè quando il regime è *all’incirca* quello di saturazione o interdizione. Il regime attivo è invece quello che si trova ben distante, nel grafico, rispetto a interdizione e saturazione.

In ogni caso, è evidente che il funzionamento come switch di un transistor, a cui abbiamo già fatto cenno in precedenza come a una delle applicazioni di rilievo, è figlio della possibilità di passare da interdizione (switch “spento”, in genere) a saturazione (switch “acceso”, in genere) agendo semplicemente sulla I_B . Facendo riferimento alla retta di carico disegnata in Fig. 4, se si suppone di far passare I_B da un valore prossimo a 0 fino a $100 \mu\text{A}$, la corrispondente I_C schizza da circa 0 a circa 100 mA : dunque una variazione piccola (in termini assoluti) della corrente di base determina una variazione decisamente grossa (in termini assoluti, e comparativi) della corrente di collettore, a cui corrisponde una diminuzione (da circa V_0 a circa 0) di V_{CE} .

VI. COMPORTAMENTO A PICCOLI SEGNALI (IN CONFIGURAZIONE A EMETTITORE COMUNE)

Un transistor in regime attivo può comportarsi come *amplificatore*. Per usare il transistor come amplificatore ci si deve preoccupare di fornire le giuste tensioni di *polarizzazione* al dispositivo. In altre parole, il dispositivo compie la sua funzione attiva solo se è “correttamente alimentato”, ovvero “polarizzato” in modo da trovarsi nel regime attivo. Quindi ogni circuito amplificatore dovrà provvedere delle *tensioni continue* che realizzino le polarizzazioni richieste. Focalizzando la nostra attenzione sulla configurazione a emettitore comune, questo vuol dire che dovremo in ogni caso preoccuparci di avere le tensioni V_{BE} e V_{CE} richieste per accedere al regime attivo, e di conseguenza avere anche le correnti I_C e I_B , entrambi *continue*, che scorrono nel transistor.

Dovrebbero ora risultare chiari alcuni punti che sono tipici quando si impiega un amplificatore a transistor:

- quello che normalmente si vuole amplificare non è la tensione, o corrente, continua di polarizzazione, ma si vuole invece che l’amplificazione avvenga a carico di *segnali alternati* dipendenti dal tempo, eventualmente sovrapposti a quelli di polarizzazione;
- di conseguenza, in un circuito amplificatore a transistor è sempre presente una parte che serve per creare le giuste tensioni (e correnti) di polarizzazione, cioè un *circuito di polarizzazione*;
- tenendo conto del fatto che segnali variabili nel tempo da amplificare e tensioni di polarizzazione sono, almeno in parte, sovrapposti, e che è necessario che il transistor operi sempre nel regime attivo, è ovvio che esistono dei limiti nell’ampiezza del segnale che può essere amplificato, quello che va all’ingresso dell’amplificatore: un’ampiezza eccessiva potrebbe infatti portare il transistor fuori dal regime attivo. Di conseguenza si parla spesso di amplificazione di *piccoli* (o *deboli*) segnali;
- infine, c’è una conseguenza “tipografica” del fatto che ci si riferisce a piccoli segnali: le grandezze che caratterizzano il comportamento in queste condizioni sono in genere scritte in *caratteri minuscoli*, come già visto per la determinazione della resistenza dinamica del diodo.

La configurazione a emettitore comune si presta ad amplificare piccoli segnali oscillanti e alternati che vengono inviati alla base. Quindi la base funge da *ingresso* per tali segnali (potenziali riferiti alla linea dell’emettitore) e l’*uscita* si ritrova al collettore (anche qui potenziali riferiti alla linea dell’emettitore, che pertanto viene normalmente collegato a terra).

La risposta ai piccoli segnali oscillanti può essere in linea di principio diversa rispetto al comportamento in

continua trattato finora. La caratteristica di amplificazione in corrente, che si indica con β_f , o anche con h_{fe} (notate i caratteri minuscoli dei pedici), è definita come segue:

$$\beta_f = \frac{\partial i_c}{\partial i_b} \Big|_{V_{CE}} , \quad (13)$$

dove i_c e i_b sono le ampiezze, o ampiezze picco-picco, delle correnti (oscillanti) che scorrono in collettore e base a causa della presenza del piccolo segnale oscillante in ingresso, e il simbolo $|_{V_{CE}}$ indica che il valore considerato si riferisce a un determinato valore della tensione V_{CE} (continua), cioè a *determinate condizioni di lavoro*. Tipicamente, anche β_f vale alcune centinaia, cioè è dello stesso ordine di grandezza di β_F , almeno nei casi di interesse pratico. Analogamente a β_F , inoltre, anch’esso dipende dalle condizioni di lavoro e soffre di una rilevante tolleranza di fabbricazione.

Misurare β_f è generalmente difficile, sia per la necessità di misurare deboli segnali di ampiezza che può facilmente confondersi con il rumore, che per l’esigenza di usare modelli interpretativi, spesso approssimati. È possibile stimare la grandezza facendo riferimento a misure in continua e definendo $\beta_f \approx \Delta I_C / \Delta I_B$, dove i simboli Δ indicano differenze tra valori continui ottenuti con polarizzazioni leggermente diverse, cioè per valori di I_B che si trovano abbastanza vicini gli uni agli altri. Un modo più raffinato, ma concettualmente simile, per determinare β_f consiste nel ripetere tante volte la misura di I_C a valori di I_B prossimi tra loro, e quindi nel graficare le coppie I_C , I_B ottenute; il guadagno in corrente per piccoli segnali alternati si può allora determinare facendo un best-fit dei dati a una retta che passa per l’origine.

Per completezza vale la pena di sottolineare come l’analisi del comportamento di un transistor per piccoli segnali oscillanti possa essere realizzata in una forma più completa e spesso più accurata utilizzando un approccio generale, detto talvolta *modello lineare di risposta*, in cui si fa uso di matrici i cui elementi descrivono tutte le proprietà di interesse (h_{fe} è uno di questi elementi). In questa sede non trattiamo tale approccio, che probabilmente incontrerete nel vostro futuro.

A. Amplificatore di tensione

Un transistor, per esempio montato in configurazione a emettitore comune e operante in regime attivo, è in grado di amplificare delle piccole correnti oscillanti con un guadagno β_f . L’interesse principale della configurazione a emettitore comune è che essa può servire anche da *amplificatore di tensione*, sempre per piccoli segnali oscillanti. A questo scopo il segnale alternato viene inviato alla base, cioè sovrapposto alla tensione continua V_{BE} , per cui l’ingresso è tra la base e terra, come rappresentato in Fig. 5(b). Notate che nello schema è stato inserito un condensatore C , che serve a “disaccoppiare” l’ingresso

rispetto alla tensione continua di polarizzazione. Infatti si vuole che questa tensione serva a polarizzare la giunzione BE, e quindi occorre evitare che fluisca corrente nel circuito che produce il segnale oscillante: il condensatore blocca le correnti continue e fa passare le componenti alternate, dato che la sua impedenza, di modulo $1/(\omega C)$, può essere ritenuta trascurabile per ω sufficientemente alto e estremamente alta per ω tendente a zero. L'uscita dell'amplificatore, invece, è tra collettore e terra.

Dette v_{in} e v_{out} le *ampiezze* (o ampiezze picco-picco) dei segnali in ingresso e in uscita, si definisce il *guadagno in tensione* come $A_v = v_{out}/v_{in}$. Facendo riferimento alle curve caratteristiche di Fig. 4, si può intuire come l'applicazione del segnale oscillante v_{in} induca una modulazione della corrente di base, cioè una modulazione oscillante i_b sovrapposta alla I_B continua di polarizzazione che determina il punto di lavoro. La modulazione fa sì che il transistor sposti (in modo oscillante) il proprio punto di lavoro: di conseguenza appare una modulazione nella corrente di collettore, che, almeno in prima approssimazione (in approssimazione *lineare*) ha ampiezza (o ampiezza picco-picco) $i_c = \beta_f i_b$.

Cerchiamo ora un collegamento tra queste correnti oscillanti e le tensioni v_{in} e v_{out} . Il segnale v_{in} è applicato alla serie della resistenza di base R_B e della *giunzione* base-emettitore. Questa giunzione è polarizzata direttamente, per cui essa viene vista come una resistenza finita. Dato che il comportamento che stiamo esaminando riguarda piccoli segnali oscillanti, la resistenza che la giunzione base-emettitore oppone a tali segnali può essere approssimata con la *resistenza dinamica* r_b della giunzione stessa, determinata per il punto di lavoro che corrisponde alla I_B continua di polarizzazione.

Quando abbiamo studiato la resistenza dinamica della giunzione di un diodo bipolare, abbiamo approssimato al primo ordine la resistenza dinamica come $r_d = (\partial I / \partial V)_{I=I_q}^{-1}$ e abbiamo trovato che essa, per una giunzione che segue l'equazione di Shockley, può essere espressa come $r_d \simeq \eta V_T / I_q$, essendo I_q la corrente di lavoro, V_T la d.d.p. legata all'energia termica ($V_T \simeq 26$ mV a temperatura ambiente) e η un parametro costruttivo della giunzione. Lo stesso approccio può essere applicato in questo caso, dato che la giunzione BE polarizzata direttamente è in questo ambito molto simile a quella di un diodo. Questo ci permette di approssimare la resistenza dinamica della base come $r_b = \eta V_T / I_B$.

Come già accennato, è necessario tenere conto che, se per un diodo al silicio il parametro η è generalmente ben approssimato da $\eta \simeq 2$, nel caso dei transistor e a causa dei dettagli costruttivi della giunzione BE (drogaggio asimmetrico, spessore sottile), lo stesso parametro assume valori più bassi, spesso prossimi all'unità ($\eta \simeq 1$). Nell'ambito dell'approccio matriciale menzionato prima, del quale non facciamo uso a causa del suo grado di complicazione, il parametro r_b è indicato dal simbolo h_{ie} , che si trova tabulato per alcuni valori di operazione nei datasheet dei transistor. Nel caso del transistor 2N1711 il parametro η dedotto dal valore tabulato di h_{ie} è pros-

simo a 1.2, ma, a causa della grande tolleranza con cui le caratteristiche del componente possono essere definite, è possibile che esso assuma valori effettivi compresi (almeno) tra 1 e 1.5.

In ogni caso, tenendo conto del collegamento in serie tra le resistenze R_B e r_b , possiamo scrivere la relazione, da intendersi come approssimata, $v_{in} = (R_B + r_b)i_b \simeq (R_B + \eta V_T / I_B)i_b$.

Per quanto riguarda il legame fra v_{out} e i_c possiamo fare riferimento all'equazione della retta di carico (Eq. 12), notando però che in questo caso siamo interessati solo ai segnali oscillanti, ovvero possiamo trascurare le tensioni continue (che invece servono per la polarizzazione del transistor). Se nell'Eq. 12 indichiamo con v_{out} l'ampiezza, o ampiezza picco-picco, della componente oscillante del segnale al collettore (legata alla corrente oscillante i_c), togliendo per il motivo appena detto il termine V_0 , che è ovviamente costante, otteniamo $v_{out} = -R_C i_c$. Il segno meno di questa relazione può essere compreso anche notando che, quando la corrente di collettore aumenta, allora c'è una maggiore caduta di tensione attraverso la resistenza R_C , e quindi la tensione v_{out} diminuisce.

Mettendo tutto insieme possiamo scrivere:

$$A_v = \frac{v_{out}}{v_{in}} = -\frac{R_C}{R_B + r_b} \frac{i_c}{i_b} \simeq -\frac{R_C}{R_B + \eta V_T / I_B} \beta_f . \quad (14)$$

Ritroviamo ancora un segno meno, che sta a significare che il segnale oscillante in uscita è *sfasato* (di $\pm\pi$) rispetto a quello in ingresso. Questa equazione dimostra che ci può sicuramente essere amplificazione di tensione ($A_v > 1$) e che il guadagno dipende non solo da β_f , ma anche dal rapporto $R_C/(R_B + r_b)$.

Notate che, in certe condizioni, in particolare per alte correnti di lavoro I_B , r_b può diventare trascurabile rispetto a R_B e pertanto $A_v \simeq -(R_C/R_B)\beta_f$. Nel caso dell'esercitazione svolta in laboratorio questa approssimazione non è generalmente valida. Nelle misure da me effettuate, la polarizzazione era tale da produrre $I_B \approx 6.2$ μ A, a cui corrispondeva, supponendo $\eta = 1.2$, $r_b \approx 5.0$ kohm. Poiché nel mio caso era $\beta_f \simeq 208$, tenendo conto che $R_B = 0.56$ kohm e $R_C = 1.0$ kohm (valori nominali, qui usati per una stima), si otteneva un valore atteso del guadagno in tensione $A_{v,att} \approx 37$. Dalle misure delle ampiezze (picco-picco) dei segnali in ingresso e in uscita avevo $v_{in} \approx 12$ mV e $v_{out} \approx 540$ mV, per un guadagno misurato $A_v \approx 45$.

Infine, osservate che la presenza di un guadagno in corrente e di un guadagno in tensione implica un *guadagno di potenza*, cioè la configurazione esaminata si comporta anche da amplificatore di potenza. È facile verificare che il guadagno in potenza, proporzionale al prodotto $\beta_f A_v$, va con il quadrato di β_f . Una stupida osservazione conclusiva: naturalmente questa amplificazione di potenza non è magicamente creata dal transistor, ma il transistor permette di sfruttare in modo opportuno la potenza resa disponibile dalle sorgenti di alimentazione, cioè nel nostro caso dal generatore V_0 , allo scopo di ottenere l'amplificazione richiesta. Inoltre per ogni transistor reale esiste un

limite alla massima potenza che può essere “manipolata” dal componente, che non può eccedere un certo limite pena la possibile rottura a causa dell’incremento di temperatura conseguente alla dissipazione. Per il transistor 2N1711 questo limite è dell’ordine di 0.8 W.

B. Resistenze di ingresso e uscita

Nelle nostre esercitazioni pratiche in genere utilizziamo segnali “di ingresso” prodotti dal generatore di funzioni e leggiamo l’“uscita” con l’oscilloscopio. In termini generali, questi strumenti hanno il primo una resistenza interna relativamente bassa ($r_G = 50$ ohm), che possiamo interpretare qui come resistenza di uscita del generatore stesso secondo l’approccio di Thévenin, e il secondo una resistenza di ingresso relativamente alta ($r_{osc} = 1$ Mohm). Questa circostanza ci permette spesso di disinteressarci dei possibili effetti dei circuiti che introduciamo tra “ingresso” e “uscita”. Infatti, parlando in modo grossolanamente, la bassa resistenza del generatore limita la possibilità che ci siano rilevanti cadute di potenziale su r_G , mentre l’alta resistenza dell’oscilloscopio evita che la presenza dello strumento “sottragga” un’eccessiva quantità di corrente dal nostro circuito.

Non sempre la situazione è così semplice. Per esempio, si verifica spesso che un circuito sia costituito da tanti sotto-circuiti collegati tra loro in serie, come abbiamo visto per esempio nel caso del circuito integratore-derivatore in cascata. In situazioni di questo tipo diventa importante poter determinare a priori le resistenze, o, meglio, impedanze, di ingresso e di uscita di ogni sotto-circuito. Con resistenza, o impedenza, di ingresso e di uscita intendiamo, in sostanza, le resistenze, o impedanze, che sono “viste” rispettivamente all’ingresso e all’uscita di ogni sotto-circuito. In quanto segue, e per pura semplicità, tratteremo di resistenze e non di impedanze, supponendo di poter trascurare gli effetti capacitivi, o, più in generale, reattivi, che si verificano negli elementi di circuito considerati. Questa affermazione è naturalmente tanto più accurata quanto più bassa è la frequenza di operazione del circuito.

La valutazione accurata delle resistenze in ingresso e in uscita a uno stadio amplificatore è non sempre banale: nella nostra descrizione ci accontenteremo di un modello molto semplificato, che trascura alcuni dettagli funzionali del transistor (essi sono considerati nel modello lineare a piccoli segnali che potete trovare in numerose fonti) e non si pone alcun problema relativo alla separazione, pratica e concettuale, fra le caratteristiche intrinseche dell’amplificatore e quelle più propriamente pertinenti alle parti di circuito destinate a produrre il segnale in ingresso e al carico in uscita. In particolare supponiamo ideale il generatore di d.d.p. v_{in} e trascurabile la corrente variabile nel tempo che potrebbe scorrere nella maglia che contiene R_P (questa resistenza è generalmente grande, per cui l’approssimazione è ragionevole).

Con queste approssimazioni, per un amplificatore realizzato con un transistor *BJT* in configurazione di emettitore comune [Fig. 5(b)] la resistenza di ingresso è la serie $R_{in} \sim R_B + r_b$ che abbiamo individuato prima. Poichè il guadagno in tensione, $A_v = R_{out}\beta_f/R_{in}$, scala come $1/R_{in}$, si cerca spesso di operare in condizioni in cui R_{in} è piccolo. Di conseguenza, la resistenza di ingresso dell’amplificatore a emettitore comune viene in genere classificata come *bassa*, grazie a un’opportuna scelta di R_B e della polarizzazione diretta della giunzione BE, che implica r_b piccola.

Come abbiamo già stabilito, per la configurazione a emettitore comune si ha $v_{out} = -R_C i_{out}$, che in pratica stabilisce $R_{out} = R_C$ (naturalmente si prende il valore positivo omettendo il segno). Possiamo giungere a questa conclusione usando un ragionamento diverso da quello impiegato prima, che era basato sull’equazione della retta di carico. Qui possiamo partire dalla constatazione che la resistenza di uscita R_{out} è in pratica la resistenza “vista” tra collettore ed emettitore. Nel circuito di Fig. 5(b), essa può essere schematizzata come la serie di resistenze delle due giunzioni, BC e BE, con in parallelo la serie di R_C e della resistenza interna r del generatore V_0 . Poichè $R_{out} = R_C$, il ramo che passa per il transistor, cioè la serie delle due giunzioni BC e BE, ha resistenza molto più alta dell’altro. Entro i limiti del dimensionamento del circuito (in particolare, la definizione del punto di lavoro del transistor), in genere esiste una certa libertà nello scegliere la resistenza di collettore e R_C viene spesso presa come piuttosto alta (valori tipici del kohm, o delle decine di kohm) così da consentire un consistente guadagno in tensione. Pertanto la resistenza di uscita dell’amplificatore a emettitore comune viene in genere classificata come *medio-alta*.

Una conseguenza del fatto che le resistenze di ingresso e uscita siano (classificate come) rispettivamente bassa e alta è evidente: se decidessimo di realizzare un amplificatore “pluri-stadio” usando diversi amplificatori a emettitore comune posti uno in seguito all’altro (in cascata, o “in serie” fra loro), lo stadio successivo potrebbe “sovrafficare” quello precedente. Infatti, a causa della richiesta di corrente dovuta alla bassa resistenza di ingresso dello stadio successivo, si potrebbe avere una rilevante caduta di potenziale nella resistenza di uscita dello stadio precedente. Il risultato sarebbe un guadagno complessivo dell’intero circuito nettamente minore delle aspettative, cioè $|A_{v,tot}| \ll \prod |A_{v,i}|$, dove $A_{v,i}$ è il guadagno atteso dello stadio i -esimo, cioè il guadagno che lo stadio avrebbe se fosse collegato a un carico altissimo, e la produttoria è estesa a tutti gli stadi. Per risolvere il problema si possono impiegare configurazioni diverse, che, secondo quanto sarà accennato in seguito, presentano resistenze di ingresso e di uscita differenti da quelle della configurazione a emettitore comune.

C. Stabilità della polarizzazione

Nei voluminosi tomì studiati dai tecnici che progettano circuiti ci sono ampi capitoli dedicati alle tecniche di polarizzazione, cioè alla discussione dei metodi e delle relative configurazioni circuituali che meglio permettono di determinare la polarizzazione del transistor. Qui non intendiamo entrare nei dettagli ma, limitandoci solo alla configurazione a emettitore comune rappresentata in Fig. 5(b), possiamo osservare alcuni aspetti potenzialmente critici.

Il problema della stabilità citato nel titolo di questa sezione nasce soprattutto dalla circostanza che il transistor può surriscaldarsi durante il suo impiego. Questa è una conseguenza della dissipazione dovuta alle correnti che lo attraversano. Infatti molto spesso i transistor sono montati su appositi dissipatori, che servono per aumentare lo scambio termico con l'ambiente e quindi ridurre la temperatura di operazione.

La dissipazione ha luogo nel transistor perché al suo interno le correnti incontrano delle resistenze. In effetti dentro un transistor ci sono sicuramente delle resistenze. In primo luogo la giunzione, se polarizzata inversamente, oppone una certa resistenza al passaggio di corrente (nell'ambito che qui stiamo esaminando, la resistenza non è quella dinamica, ma quella "effettiva": infatti ci stiamo occupando della polarizzazione delle giunzioni). Se la giunzione è polarizzata direttamente, questa resistenza cala, fino a diventare virtualmente trascurabile: tuttavia, prima di arrivare alla giunzione, i portatori di carica devono attraversare uno spessore non necessariamente trascurabile (in particolare nel collettore e nell'emettitore) di materiale semiconduttore drogato. Questo materiale ha una certa resistività, o mobilità, e dunque, in serie alla resistenza (effettiva) della giunzione propriamente detta dobbiamo considerare delle resistenze addizionali, così come fatto nel diodo.

I materiali semiconduttori hanno una caratteristica di temperatura di tipo *NTC*, cioè la loro resistività diminuisce con l'aumentare della temperatura; quindi le resistenze in serie di cui stiamo parlando tendono a diminuire di valore quando il transistor si surriscalda.

Vediamo una prima conseguenza che riguarda la polarizzazione della giunzione BE: Nello schema di Fig. 5(b) essa è determinata dalla V_{BE} , ottenuta a sua volta partendo da V_0 attraverso la caduta di potenziale nella serie $R_P + R_B$. L'entità della caduta di potenziale dipende linearmente da I_B che, ancora a sua volta, dipende in maniera molto accentuata (se la polarizzazione è diretta, come stiamo considerando qui) da V_{BE} [vedi Fig. 4(a)]. Se il transistor si surriscalda, la resistenza effettiva della giunzione BE può diminuire, la corrente I_B può aumentare di intensità e così anche la caduta di potenziale sulla serie $R_P + R_B$, determinando una variazione del punto di lavoro della giunzione BE rispetto alle specifiche di progetto.

Per limitare l'effetto, spesso al posto della resistenza R_P si utilizza un partitore di tensione, rappresentato in

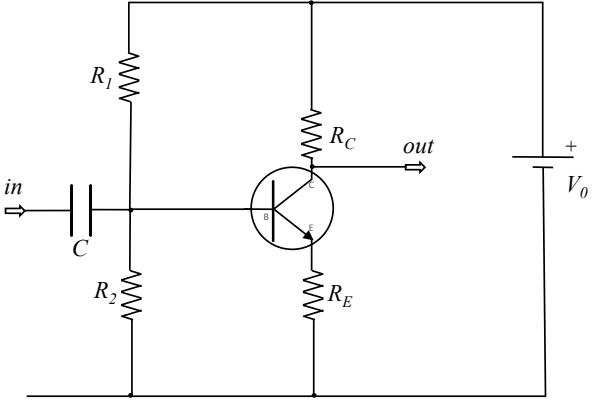


Figura 6. Esempio di possibile circuito di polarizzazione del transistor a emettitore comune, con disegnati i resistori di cui si discute nel testo.

Fig. 6 dalle resistenze R_1 e R_2 . Inoltre, in genere R_B viene scelta molto minore rispetto a R_1 e R_2 , anzi, spesso R_B è proprio assente (essa non c'è nello schema di Fig. 6). Dunque la resistenza effettiva della giunzione viene a trovarsi in parallelo a R_2 e, dimensionando in maniera opportuna R_2 , si può fare in modo che i suoi effetti, e quelli conseguenti alla sua eventuale variazione con la temperatura, siano trascurabili. In altre parole, il dimensionamento del partitore è scelto in modo che l'intensità della corrente che scorre nella serie $R_1 + R_2$ sia molto maggiore rispetto a I_B , così che la d.d.p. V_{BE} sia praticamente indipendente da I_B e dalle sue eventuali variazioni con la temperatura.

È utile osservare che la presenza del partitore di tensione all'ingresso dell'amplificatore, oltre a rendere meglio definito il punto di lavoro della giunzione BE, ha anche una conseguenza importante sulla resistenza di ingresso. Infatti per il circuito di Fig. 6 la resistenza di ingresso è data dal parallelo tra la serie $R_B + r_b$ e R_2 : per evitare di ridurre ulteriormente la già piccola resistenza di ingresso, in genere il partitore viene dimensionato evitando l'impiego di valori eccessivamente piccoli per R_2 rispetto a r_b .

D. Resistenza di emettitore

Passiamo ora ad esaminare le conseguenze che surriscaldamento e variazione della resistenza effettiva della giunzione possono avere nel circuito di collettore. Qui i problemi sono accentuati dall'intensità della corrente I_C e dal potenzialmente elevato valore della resistenza effettiva offerta dalla giunzione BC quando essa è polarizzata inversamente. La combinazione di queste due circostanze determina una rilevante dissipazione di potenza a carico della giunzione BC, e dunque del componente.

Osserviamo che, se la resistenza effettiva della giunzione diminuisse in presenza di surriscaldamento, la corrente di collettore tenderebbe ad aumentare. Infatti la retta di

carico corrispondente diventerebbe più inclinata, e quindi il punto di lavoro si sposterebbe verso correnti I_C di intensità maggiore. Questo potrebbe a sua volta indurre un aumento della potenza dissipata, e quindi un ulteriore surriscaldamento, innescando un meccanismo a catena il cui esito potrebbe essere letale per il transistor (al di sopra di una certa temperatura la giunzione si danneggia irreparabilmente).

Per prevenire questa possibilità è in genere sufficiente aggiungere un resistore, denominato R_E in Fig. 6, tra emettitore e linea di massa, o di terra. La corrente che fluisce in questo resistore ha un'intensità pressoché identica a quella che fluisce nel collettore ($|I_E| = |I_B| + |I_C| \simeq |I_C|$ per l'effetto transistor). Essa provoca quindi una caduta di potenziale non trascurabile pur in presenza di piccoli valori di R_E , in conseguenza della quale l'emettitore viene a trovarsi a una d.d.p. diversa da zero (e positiva) rispetto alla linea di massa, o di terra. La d.d.p. V_{BE} assume allora un valore *minore* rispetto a quello che si misurerrebbe in assenza di R_E ; questo comporta una diminuzione della corrente I_B , il cui effetto principale è quello di ridurre l'intensità della corrente I_C . Dunque la presenza di R_E può effettivamente mitigare l'effetto potenzialmente disastroso del surriscaldamento.

Semplificando e cercando di generalizzare, la presenza della resistenza R_E nel circuito di polarizzazione che abbiamo appena descritto si traduce nel meccanismo seguente: I_C aumenta a causa dell'aumento di temperatura, dunque aumenta I_E e di conseguenza la caduta di potenziale ai capi di R_E . Questo provoca una riduzione di V_{BE} , che comporta una diminuzione di I_C e quindi degli effetti del surriscaldamento. Allora, usando un linguaggio molto descrittivo, potremmo dire che R_E gioca un ruolo che consiste nel monitorare quello che succede (quanto vale I_C) e impartire le corrette istruzioni al sistema affinché esso tenda a ri-stabilizzarsi alle sue condizioni di operazione ordinarie.

Quello qui ipotizzato è un possibile prototipo di meccanismo di *feedback*, che in italiano si traduce per esempio con *retroazione*: la retroazione conseguente all'aumento di I_C al di là del valore di progetto conduce, attraverso un meccanismo “indiretto”, a una diminuzione di I_C .

Parlare di feedback in presenza di un amplificatore apre un interessantissimo scenario su cui torneremo in dettaglio con un'altra nota. Qui ci limitiamo a valutare le conseguenze che la presenza di R_E comporta nell'amplificazione di piccoli segnali oscillanti, in particolare nel guadagno. Analogamente a quanto si verifica per le correnti e tensioni (continue) di polarizzazione, ai capi di questa resistenza si produce una d.d.p. (oscillante) di ampiezza, o ampiezza picco-picco, pari a $R_E i_e$, dove i_e è l'ampiezza, o ampiezza picco-picco, della corrente (oscillante) che passa per la resistenza fuoriuscendo dall'emettitore. Tale intensità di corrente è, in modulo, pressoché pari a i_c a causa dell'effetto transistor. Più correttamente, è $i_e = i_c / \alpha_f$, dove con α_f indichiamo il coefficiente dell'effetto transistor rilevante per piccoli segnali variabili nel tempo. Osservate che, solo quando si considera

il transistor come nodo di corrente, occorre aggiungere un segno, come fatto in Sez. III. Infatti le due correnti di emettitore e di base hanno segni opposti rispetto al nodo costituito dal transistor, e segni evidentemente concordi nell'equazione di maglia che qui ci interessa (per esempio, quando la corrente entra nel collettore, cioè nelle fasi in cui il segnale di ingresso è tale da provocare questa situazione, essa esce dall'emettitore mantenendo lo stesso verso di percorrenza della maglia). Quindi la d.d.p. oscillante ai capi della resistenza di emettitore è $R_E i_c / \alpha_f = R_E i_b \beta_f / \alpha_f$, dove $\alpha_f / \beta_f = 1 + \beta_f$. Esaminando la maglia di base, nella quale scorre la corrente di intensità con ampiezza i_b , si ha

$$v_{in} = (R_B + r_b + (1 + \beta_f)R_E)i_b, \quad (15)$$

mentre è sempre

$$v_{out} = -R_C i_c, \quad (16)$$

da cui

$$A_{V,RE} = \frac{v_{out}}{v_{in}} = -\beta_f \frac{R_C}{R_B + r_b + R_E(1 + \beta_f)} \simeq -\frac{R_C}{R_E}, \quad (17)$$

dove l'ultimo passaggio costituisce un'ulteriore approssimazione, valida quando $R_E \beta_f \gg R_B + r_b$.

In definitiva, allora, aggiungere una resistenza all'emettitore porta a *diminuire* il guadagno in tensione dell'amplificatore di tensione a emettitore comune. Questa diminuzione porta a indicare come *negativo* il feedback esercitato da tale resistenza. Nel mio caso, usando $R_E = 68$ ohm nominali, ho misurato, nelle stesse condizioni descritte prima, un guadagno $A_{v,RE} \approx 10$ a fronte di una stima di valore atteso, determinata dall'Eq. ??, di 11.

Per mitigare la diminuzione di guadagno mantenendo i vantaggi del feedback in termini di stabilità di polarizzazione (cioè quelli relativi alle d.d.p. e correnti continue di polarizzazione) è possibile ricorrere a un “trucco” che consiste nel montare un condensatore C_E in parallelo a R_E . In queste condizioni, la resistenza di emettitore viene vista per intero solo in condizioni stazionarie, cioè quelle che riguardano la polarizzazione. All'aumentare della frequenza l'impedenza $|Z_{CE}| = 1/(\omega C_E)$ tende a zero, per cui anche la d.d.p. ai capi del parallelo tende ad annullarsi, ripristinando le condizioni di guadagno viste per l'amplificatore a emettitore comune senza la resistenza di emettitore, con un comportamento che segue grossolanamente l'andamento di un filtro passa-alto.

Oltre a prevenire il surriscaldamento, la resistenza di emettitore, ovvero una impedenza di modulo non trascurabile tra emettitore e linea di massa, o terra, ha anche un ulteriore aspetto benefico. Uno dei principali limiti nell'operazione di un amplificatore a emettitore comune è rappresentato dalla *distorsione di non linearità*, che comporta una modifica della forma d'onda in uscita rispetto a quella in ingresso. Gli effetti di distorsione possono diventare evidenti quando l'ampiezza del segnale v_{in}

(supposto sinusoidale) supera un certo valore: la forma di v_{out} si discosta sempre più da quella di una sinusoidale. L'effetto è ancor più evidente impiegando una forma d'onda triangolare, dove la distorsione è ben visibile nella modifica della forma dei tratti, idealmente rettilinei, del triangolo. La distorsione è dovuta al fatto che, non appena v_{in} si avvicina, in ampiezza, a V_T (cioè, nella pratica, per v_{in} dell'ordine di poche decine di mV), la corrente di base i_b tende ad assumere un andamento fortemente non lineare nei confronti di v_{in} . Inoltre, poiché la corrente di base varia dinamicamente secondo la $i_b \sim v_{in}/R_{in}$, un segnale in ingresso di ampiezza rilevante può addirittura condurre (dinamicamente) il transistor fuori dal regime attivo. Qui, non vale più la relazione lineare tra correnti di base e di collettore, da cui la distorsione. Tenendo conto della curva caratteristica di ingresso [Fig. 3(a)], è evidente che la corrente di base varia più marcatamente in corrispondenza della semionda positiva del segnale in ingresso, dove la corrente di base può facilmente diventare così alta da portare il transistor vicino al regime di saturazione. Normalmente, infatti, la distorsione è “asimmetrica”, e le semionde negative del segnale in uscita vengono “tosate” (spianate, in inglese si parla di “clipping”).

Il fenomeno della distorsione produce una sensibile limitazione al *range dinamico* accettato in ingresso dall'amplificatore, cioè all'intervallo di ampiezze v_{in} per le quali l'amplificatore funziona in modo lineare. Il feedback negativo, riducendo i_b a parità di v_{in} , comporta un aumento del range dinamico, e quindi la possibilità di operare in regime lineare per una più estesa gamma di segnali in ingresso. Nell'esercitazione pratica da me svolta, operando senza resistenza di emettitore gli effetti di distorsione diventavano chiaramente evidenti già per $v_{in} \sim 30$ mV, mentre la presenza della resistenza di emettitore permetteva di estendere l'ampiezza in ingresso fino a diverse centinaia di mV.

VII. ALTRE CONFIGURAZIONI

Benché quella a emettitore comune sia probabilmente la configurazione più frequente in cui vengono utilizzati i transistor *BJT*, vale la pena di fare un breve cenno alle configurazioni *a base comune* e *a collettore comune*, detta anche *emitter follower* nel caso di transistor *npn*. In questo cenno non ci preoccupiamo dei dettagli circuitali che permettono, ad esempio, di realizzare la corretta polarizzazione delle giunzioni, che deve essere sempre tale da portare il transistor nel regime attivo. Inoltre supporremo, come ovvio, di analizzare il comportamento del transistor come amplificatore, determinando qualitativamente resistenze di ingresso e uscita e guadagno.

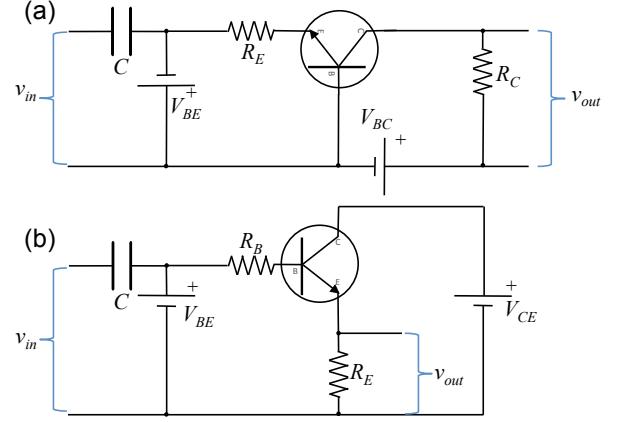


Figura 7. Schemi circuitali di possibili configurazioni con transistor *npn* a base comune (a) e emitter follower (b).

A. Base comune

La configurazione a base comune l'abbiamo, in realtà, già trattata quando abbiamo discusso l'effetto transistor [vedi Fig. 2(a)]. Nella Fig. 7(a) mostriamo un possibile schema circuitale che corrisponde alla configurazione a base comune: l'ingresso del (debole e oscillante) segnale da amplificare è inviato tra emettitore e base, però l'uscita è presa ai capi della resistenza indicata con R_C , per cui potrebbe sembrare che il terminale di base non sia a comune tra ingresso e uscita. In realtà, poiché siamo interessati al comportamento rispetto a segnali variabili nel tempo e possiamo supporre che il generatore di d.d.p. continua V_{BE} sia ideale, dunque con resistenza interna trascurabile, lo schema proposto è tecnicamente compatibile con la definizione di base comune.

Per prima cosa, osserviamo che in questa configurazione *non c'è* guadagno in corrente, che, qui, è dato dal rapporto i_c/i_e . Infatti $I_C = \alpha_F I_E \lesssim I_E$, e da qui possiamo dedurre che anche $i_c \lesssim i_b$.

Determiniamo ora, sempre in modo molto approssimativo e senza entrare nei dettagli, le resistenze di ingresso e di uscita. All'ingresso il segnale vede la serie $R_{in,bc} \sim R_E + r_b$, dove r_b può essere piccola a causa della polarizzazione diretta della giunzione BE. In uscita, invece, la resistenza è data dal parallelo tra R_C e la resistenza della giunzione BC, polarizzata inversamente e dunque dotata di un'elevata resistenza dinamica che qui indichiamo con r_c . Dunque, supponendo r_b trascurabile, $R_{in,bc} \sim R_E$ e $R_{out,bc} \sim R_C // r_c$. Approssimando come unitario il guadagno in corrente, il guadagno in tensione diventa allora $A_{v,bc} = v_{out}/v_{in} = R_{out,bc}/R_{in,bc}$, dove, come si vede facilmente, ingresso e uscita sono in fase tra loro. Sulla base delle considerazioni approssimative che abbiamo svolto, per avere amplificazione di tensione si può scegliere $R_C >> R_E$, in modo tale che, a prescindere dalle resistenze dinamiche delle giunzioni, si abbia strettamente $A_{v,bc} \sim R_C/R_E > 1$. A livello di classificazione, si dice spesso che la configurazione a base comune presenta

una resistenza di ingresso *bassa* e una resistenza di uscita *alta*, in definitiva simili a quelle della configurazione a emettitore comune.

Comparata alla configurazione a emettitore comune, la base comune non risulta avere vantaggi evidenti. A parità (in termini di classificazione) di resistenze di ingresso e di uscita, questa configurazione è meno “efficiente” a causa del guadagno in corrente che, al massimo, vale un po’ meno dell’unità. Tuttavia essa ha applicazioni tradizionali, in particolare nel settore delle alte frequenze.

B. Emitter follower (o “collettore comune”)

Una schema di configurazione emitter follower è raffigurata in Fig. 7(b). Lo schema mostrato sembra in contraddizione con la definizione di collettore comune che qualche volta si attribuisce a tale configurazione: infatti ingresso e uscita sono riferiti alla linea di massa, o di terra, che in questo caso *non* è quella del collettore. Questo è dovuto al fatto che in questa nota facciamo riferimento a transistor *npn*: si potrebbe infatti dimostrare che, se il transistor fosse *pnp*, la configurazione a collettore comune vedrebbe i segnali di ingresso e uscita riferiti alla linea del collettore, che quindi sarebbe effettivamente a comune. Per evitare eccessive discussioni, in particolare sulle polarità delle tensioni di polarizzazione, facciamo riferimento alla configurazione di Fig. 7(b), prendendo per buona l'affermazione che essa rappresenta, di fatto, una configurazione a collettore comune.

Considerando piccoli segnali alternati in ingresso, il guadagno in corrente è proporzionale a $i_e/i_b = -(i_c + i_b)/i_b = -(\beta_f + 1)$, che è dunque, in valore assoluto, maggiore dell’unità e anche maggiore (in maniera trascurabile, essendo tipicamente $\beta_f \gg 1$) rispetto a quello della configurazione a emettitore comune. Ragionando nell’ambito delle nostre consuete approssimazioni, e seguendo il ragionamento svolto prima a proposito della presenza della resistenza di emettitore R_E , la resistenza di ingresso $R_{in,cc}$ è data dalla serie $R_{in,cc} \sim R_B + r_b + R_E(1 + \beta_f)$: il valore corrispondente può essere tale da far classificare la resistenza di ingresso della configurazione a collettore comune *alta* a causa della presenza del fattore $(1 + \beta_f)$ che si trova a moltiplicare R_E .

Nel nostro modello approssimato la resistenza di uscita $R_{out,cc}$ può essere ottenuta notando che $v_{out} = R_E i_e$, per cui essa coincide con R_E . Normalmente, è possibile fare in modo che $R_{out,cc}$ assuma valori che permettono di classificarla come *bassa*, e certamente essa è minore di $R_{in,cc}$. Pertanto la configurazione a collettore comune può realizzare condizioni di resistenza, o impedenza, di ingresso e uscita che sono *diametralmente opposte* rispetto a quelle della configurazione a emettitore comune. Dunque l’impiego di questa configurazione è sicuramente ben giustificato nel caso di circuiti pluri-stadio, per eseguire un corretto matching delle resistenze dei vari stadi e limitare i problemi di “sovracarico” che abbiamo brevemente accennato in Sez. VI B.

Characteristic	Common Base	Common Emitter	Common Collector
Input Impedance	Low	Medium	High
Output Impedance	Very High	High	Low
Phase Angle	0°	180°	0°
Voltage Gain	High	Medium	Low
Current Gain	Low	Medium	High
Power Gain	Low	Very High	Medium

Figura 8. Specchietto riassuntivo delle principali caratteristiche per le diverse configurazioni di collegamento di un transistor *BJT* (tratto da <http://www.electronics-tutorials.ws>).

Il guadagno in tensione della configurazione può essere valutato come $A_{v,cc} \sim (\beta_f + 1)R_{out,cc}/R_{in,cc}$ e si può dimostrare che il suo valore tipico è circa *unitario*; infatti, normalmente $R_{out,cc} \ll R_{in,cc}$ e, di fatto, si verifica che, per dimensionamenti tipici delle resistenze del circuito, si ha $R_{out,cc}/R_{in,cc} \approx 1/\beta_f$. Dunque il segnale in uscita ha ampiezza simile (e anche lo stesso segno, cioè la stessa fase) del segnale in ingresso. Questa circostanza, unita all’impiego che tradizionalmente viene riservato a questa configurazione, ha portato storicamente alla denominazione di *inseguitore di emettitore* (emitter follower), che sottolinea come il segnale in uscita (che si trova all’emettitore, nel caso di transistor *npn*) “inseguiva” (cioè sia simile in ampiezza e fase a) quello di ingresso.

Da ultimo, la Fig. 8, tratta da <http://www.electronics-tutorials.ws>, riporta in un semplice specchietto mnemonico le principali caratteristiche delle varie configurazioni a cui abbiamo fatto cenno in questa nota.

Appendice: Porte logiche con transistor *BJT*

Come più volte anticipato in questa nota, l’enorme diffusione dei transistor nella tecnologia attuale si deve alla possibilità di implementare con essi delle *porte logiche*, cioè dispositivi in grado di manipolare segnali logici binari, tipicamente d.d.p. “acceso” o “spento” (ovvero “alte” o “basse”, ovvero ancora 1 o 0 nella logica binaria), secondo delle funzioni definite. Esistono diversi standard per stabilire a cosa corrisponde praticamente un segnale acceso o spento: lo standard TTL (Transistor-Transistor Logic), cui abbiamo già fatto cenno, è uno di questi standard, per altro ampiamente sorpassato nell’elettronica digitale. Nell’ambito dell’elettronica digitale il transistor è un componente rilevante perché può comportarsi da switch; è evidente che il transistor *BJT*, in cui correnti controllano correnti, non è il più indicato per realizzare porte logiche fortemente miniaturizzate e in grado di operare con grandissima rapidità, obiettivi in ovvio contrasto con la dissipazione di potenza per effetto Joule associata al passaggio di correnti. Infatti nella stragrande maggioranza

dei casi l'elettronica digitale fa uso di transistor a effetto di campo (per esempio, MOS-FET), in cui sono dei campi elettrici, cioè delle d.d.p., a controllare altre d.d.p., a cui possono essere associate delle (deboli) correnti.

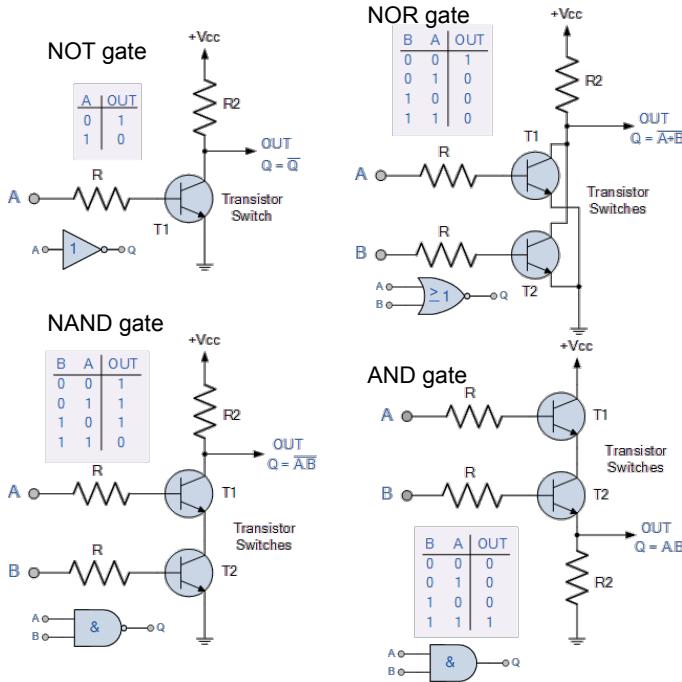


Figura 9. Alcune tra le principali porte logiche realizzate con transistor BJT: $+V_{CC}$ indica il collegamento al polo positivo del generatore di d.d.p. usato per il circuito (quello che in genere chiamiamo V_0), R e R_2 sono resistenze opportunamente dimensionate. Accanto allo schema di massima della porta è riportata la *tavola della verità* corrispondente, che mostra la manipolazione eseguita sui segnali logici (binari) applicati ai circuiti. Inoltre la figura mostra anche una possibile rappresentazione dei simboli circuituali corrispondenti alle varie porte (figure tratte da <https://www.electronics-tutorials.ws/category/logic>).

In ogni caso è esistita ed esiste tuttora la possibilità di costruire porte logiche usando transistor BJT. Alcune delle più semplici porte logiche di uso comune possono infatti essere realizzate usando gli schemi di massima (privi di dettagli e non necessariamente "ben funzionanti") riportati come esempio in Fig. 9, dove sono anche specificate le *tavole della verità* relative, cioè le corrispondenze tra valori binari in ingresso e in uscita. Per esempio, nel caso del *NOT gate* l'applicazione di un livello 1 in ingresso fa scorrere una sufficiente quantità di corrente nella giunzione BE tale da portare il transistor in regime di saturazione. In queste condizioni la corrente di collettore I_C tende al suo valore massimo (compatibilmente con le caratteristiche del transistor e con il valore della resistenza di collettore R_2) e quindi V_{CE} tende a zero. Poiché l'uscita è proprio rappresentata dalla d.d.p. fra collettore e linea di terra, cioè è V_{CE} , l'uscita stessa tende al valore 0. Se l'ingresso è invece al valore 0, il transistor va in interdizione, la corrente di collettore I_C tende a zero e l'uscita, cioè V_{CE} , tende al suo valore massimo (V_{CC} per la simbologia della figura). In altre parole, l'uscita è la negazione dell'ingresso, ovvero, usando la simbologia di figura, $Q = \bar{Q}$ (ovvero $Q = \bar{A}$, con Q uscita e A ingresso).

Nel *NOR gate* per avere un'uscita a livello 0 occorre che *almeno* uno dei due transistor, i cui collettori sono collegati assieme, sia in saturazione, cioè che *almeno* uno dei segnali in ingresso A o B sia a livello 1. Se invece entrambi gli ingressi sono a livello 0, entrambi i transistor sono in interdizione e l'uscita va a livello 1. Riassumendo in formula "logica", si ha $Q = \overline{A + B}$.

Nel *NAND gate* i due transistor sono montati in cascata e quindi è sufficiente che uno dei due abbia un uscita a livello 0, cioè che uno dei due abbia un ingresso a livello 1, perché l'uscita sia a livello 0, cioè $Q = \overline{AB}$.

Infine, l'*AND gate* lavora come un *NAND* "negato": a questo scopo è sufficiente collegare l'uscita all'emettitore della cascata dei due transistor, interponendo la resistenza R_2 verso terra, così come in un emitter follower. L'operazione logica corrispondente è quindi $Q = AB$.

Curva caratteristica di collettore del transistor con Arduino

francesco.fuso@unipi.it

(Dated: version 5 - Lara Palla e Francesco Fuso, 22 febbraio 2018)

Questa nota discute alcuni aspetti di interesse per l'esperienza di registrazione della curva caratteristica di collettore del transistor bipolare, condotta in laboratorio usando Arduino.

I. INTRODUZIONE

L'obiettivo dell'esperienza qui discussa è la ricostruzione della *famiglia di curve caratteristiche di uscita, o di collettore*, del transistor *n-p-n* 2N1711 disponibile in laboratorio. L'acquisizione delle curve, cioè delle coppie di punti I_C, V_{CE} per una data intensità di corrente di base I_B , è eseguita in maniera automatica usando Arduino. L'obiettivo è simile a quello che ci eravamo posti con l'acquisizione della curva caratteristica I vs V del diodo a giunzione, ma questa volta, per scopi prevalentemente didattici, useremo un approccio diverso, che fa uso del generatore di funzioni.

Lo scopo principale dell'esperienza è infatti la realizzazione di un sistema *sincronizzato* di acquisizione dati, una necessità molto frequente quando si richiede l'analisi di segnali generalmente dipendenti dal tempo. Il concetto e le modalità pratiche della sincronizzazione (con Arduino) erano già state introdotte all'epoca dell'acquisizione di segnali periodici, in particolare per la costruzione di record contenenti un gran numero di misure (coppie tempo, d.d.p. digitalizzata). Qui essa viene applicata a un esperimento che ha finalità specifiche. D'altra parte, anche il contenuto "fisico" dell'esperienza ha un suo valore: le tecnologie di costruzione dei transistor prevedono grosse tolleranze nei parametri di funzionamento e conoscere la caratteristica di uscita dello specifico componente che si ha a disposizione può essere molto utile. Infatti, attraverso il metodo della *retta di carico* tale conoscenza permette di stabilire il punto di lavoro effettivo del componente nelle condizioni di operazione, come verificheremo in una prossima esperienza. Inoltre qui potremo valutare il guadagno in corrente β_F (o h_{FE}) del transistor e anche stimare il guadagno β_f (o h_{fe}) per piccoli segnali oscillanti, che sono anche parametri di interesse per l'interpretazione di una futura esperienza.

II. CURVA CARATTERISTICA DI USCITA

La curva caratteristica di uscita, o di collettore, del transistor rappresenta la dipendenza della corrente di collettore I_C in funzione della tensione V_{CE} fra collettore e emettitore, misurata a una certa corrente di base, o di *polarizzazione* di base, I_B . Se I_B viene variato discretamente in un certo intervallo, allora si ricostruisce una *famiglia di curve*, ognuna per una certa I_B .

L'interesse pratico generale, e nostro in particolare, è quello di tracciare curve caratteristiche che rappresentino

il funzionamento del transistor nel *regime attivo*, o nei suoi "pressi". Questo vuol dire che

- la giunzione base-emettitore deve essere polarizzabile in modo diretto, cioè occorre che la base si trovi a potenziale positivo rispetto all'emettitore (per un transistor *n-p-n* come il 2N1711) e che questo potenziale possa arrivare al valore di soglia V_{thr} , o superarlo [1];
- è opportuno esplorare valori di I_B che siano sufficientemente piccoli (nel nostro caso fino a *poche decine di μA*) per evitare che il transistor lavori prevalentemente in *regime di saturazione*;
- di conseguenza la d.d.p. V_{BE} tra base ed emettitore dovrà essere esplorata in un range che al massimo raggiungerà, o andrà poco oltre, la tensione di soglia V_{thr} ; dal punto di vista pratico, questo si può tradurre nella richiesta $0 < V_{BE} \lesssim 0.55 - 0.75$ V.

Il transistor disponibile in laboratorio è montato su una basetta, alloggiata in un telaietto con un certo numero di boccole. Oltre al transistor, che è contattato con uno zoccolo (il montaggio è molto delicato), la basetta ospita un resistore $R_B = 560$ ohm nominali collegato in serie alla base, la cui presenza è praticamente irrilevante per i nostri scopi. Inoltre il collettore è collegato a due resistori $R_C = 1$ kohm e $R_C = 2.2$ kohm (nominali), che possono essere selezionati alternativamente, a seconda della boccola impiegata, e giocano il ruolo di *resistenza di collettore*. La basetta, che va osservata con attenzione per capirne i collegamenti, ha in definitiva lo schema rappresentato nel box tratteggiato di Fig. 1(a).

Per eseguire la misura, impieghiamo una configurazione *a emettitore comune*, in cui l'emettitore è collegato alla linea di riferimento dei potenziali, ovvero la linea di massa, o di terra. Per eseguire la misura c'è poi bisogno di prevedere due distinti sotto-circuiti, uno dedicato alla polarizzazione di base, cioè a fornire la d.d.p. V_{BE} necessaria per far lavorare il transistor attorno al regime attivo, e l'altro dedicato a fornire la d.d.p. V_{CE} (oltre che permettere la misura di I_C).

Dal punto di vista concettuale, il circuito mostrato in Fig. 1(a) soddisfa queste richieste. La parte di sinistra del circuito ipotizza un generatore di d.d.p. variabile V_{BE} e la presenza di un amperometro per monitorare il valore di I_B . Quella di destra ipotizza un generatore di d.d.p. variabile, qui chiamata V_{CC} , che controlla la d.d.p. V_{CE} e dunque la polarizzazione della giunzione

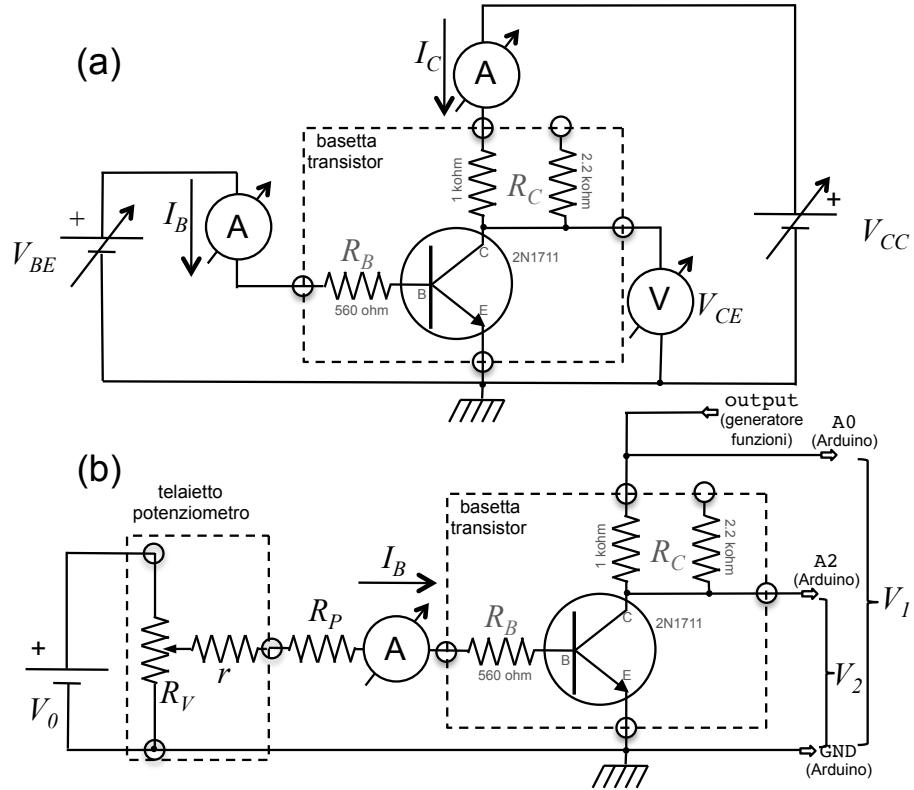


Figura 1. Schema concettuale di un’esperienza di ricostruzione della curva caratteristica I_C vs V_{CE} di un transistor *n-p-n*, del tipo di quelli disponibili in laboratorio (a); implementazione dell’esperienza di ricostruzione automatizzata della curva caratteristica mediante Arduino (b).

collettore-base. Come chiariremo in seguito, la scelta delle polarità e dello schema impiegato garantiscono che la giunzione collettore-base (o base-collettore) sia polarizzabile inversamente, come richiesto per il funzionamento del transistor nel regime attivo. Per la costruzione della curva caratteristica, si può supporre che V_{CC} venga variata, e che per ogni valore prescelto vengano misurate V_{CE} con un voltmetro e I_C con un amperometro, in modo da poter ricostruire la curva caratteristica per punti [2].

A. Polarizzazione della base

Nell’esperienza pratica si consiglia di acquisire alcune curve I_C vs V_{CE} per diversi valori di I_B fissati nel range approssimato $1 - 40 \mu\text{A}$, che corrispondono grossolanamente a valori di V_{BE} che vanno da poco sotto a poco sopra V_{thr} . In laboratorio non abbiamo disponibile un generatore di d.d.p. con le caratteristiche necessarie: esso va dunque costruito.

Lo schema rappresentato nella parte di sinistra di Fig. 1(b) rappresenta una possibile soluzione. In esso l’alimentatore $V_0 \sim 5$ V, impiegato in molte altre esperienze di laboratorio, è seguito da un partitore di tensione variabile realizzato con un potenziometro (resistore variabile) R_V , che è anche stato già impiegato in una

precedente esperienza pratica. La posizione angolare dell’alberino del potenziometro determina la resistenza tra il terminale centrale (in serie a questo si trova una piccola resistenza $r = 100$ ohm nominali, che è praticamente irrilevante per i nostri scopi) e i terminali estremi. Dunque in questo modo si realizza un partitore di tensione che ha un rapporto di partizione dipendente dalla posizione angolare dell’alberino: all’‘uscita’ di questo partitore si ottiene una d.d.p. variabile, idealmente con continuità [3], da circa 0 a circa V_0 . All’‘uscita’ del partitore si trova un’ulteriore resistenza, denominata R_P (il simbolo sta per resistenza *di polarizzazione*), che è collegata alla base (si trascura la presenza del resistore R_B). La resistenza R_P serve per limitare la corrente che fluisce verso la base ai valori di nostro interesse (al massimo poche decine di μA): si consiglia di usare $R_P = 68$ kohm (oppure $R_P = 330$ kohm) nominali (non c’è bisogno di misurare, e potete facilmente dedurre quali siano i corretti parametri di dimensionamento sulla base del range di valori desiderati per I_B). Un (micro)amperometro collegato in serie con la base permette di monitorare I_B , che dunque varierà *non linearmente* in funzione della posizione angolare dell’alberino del potenziometro. Nell’esperienza pratica si consiglia di monitorare anche V_{BE} , collegando un voltmetro ad alta impedenza di ingresso in parallelo: tuttavia questo non è strettamente necessario, visto che le condizioni di polarizzazione di base del transistor sono

completamente determinate da I_B .

La parte di destra dello schema di Fig. 1(b) mostra la strategia adottata per la variazione di V_{CC} , qui indicato come V_1 , e la misura di I_C .

Partiamo da quest'ultima: dato che vogliamo usare Arduino e che Arduino digitalizza delle tensioni analogiche in ingresso, dobbiamo trovare il modo di convertire I_C in una d.d.p.. Questo può essere facilmente ottenuto misurando la caduta di potenziale su una resistenza in serie al collettore, in pratica usando la stessa tecnica che abbiamo per esempio impiegato per ricostruire la curva caratteristica del diodo a giunzione. Questa resistenza è R_C , che è già presente nella basetta del transistor: il valore consigliato è $R_C = 1$ kohm nominali (questa resistenza va misurata con il multmetro); tenendo conto del tipico fattore di calibrazione di Arduino ($\xi \simeq 5$ mV/digit), la sensibilità nella misura di I_C risulta dell'ordine di alcuni μA , che è sicuramente sufficiente per i nostri scopi. Ovviamente la misura della caduta di potenziale su R_C richiede di usare due ingressi analogici di Arduino (i pin A0 e A2) e misurare due d.d.p. rispetto a terra, indicate con V_1 e V_2 nello schema; V_2 coincide proprio con V_{CE} , e quindi, come già realizzato nell'esperienza della curva caratteristica del diodo, la misura delle due tensioni permette di conoscere sia I_C che V_{CE} . Si ha infatti

$$V_{CE} \equiv V_2 \quad (1)$$

$$I_C \equiv \frac{V_1 - V_2}{R_C}. \quad (2)$$

Anticipiamo che, come già notato nell'esperienza della curva caratteristica del diodo, questa configurazione sperimentale ha dei limiti nei valori di V_{CE} che possono essere raggiunti. Infatti, per alte intensità di corrente I_C la caduta di potenziale su R_C può diventare così grande da impedire a V_{CE} di crescere pur in presenza di un aumento di V_1 . Dunque il metodo qui presentato non è adatto per misure di I_C superiori a un certo valore (pochi mA) e alcune delle curve acquisite potranno risultare “monche”.

L'ultimo ingrediente che dobbiamo reperire per la nostra ricetta è una sorgente variabile di d.d.p. (V_1) che giochi il ruolo della V_{CC} indicata in Fig. 1(a). Nell'esperienza della curva caratteristica del diodo abbiamo impiegato a questo scopo un'uscita PWM di Arduino seguita da un integratore RC. Potremmo implementare la stessa strategia anche in questo caso. Però, per scopi prevalentemente didattici, cioè per discutere alcuni aspetti generali e per semplificare il montaggio, in questa esperienza usiamo un altro approccio: infatti la d.d.p. variabile V_1 è qui ottenuta dal generatore di funzioni, regolato in termini di forma d'onda, ampiezza, offset e frequenza per creare un'onda triangolare di caratteristiche opportune.

B. Regolazione generatore funzioni e sincronizzazione

La filosofia della misura è a questo punto chiara: aggiustata, tramite il potenziometro, la corrente I_B a un dato

valore, il generatore di funzioni invia una d.d.p. variabile linearmente nel tempo; a intervalli regolari, separati tra di loro da una quantità Δt definibile via software, Arduino campiona le d.d.p. V_1 e V_2 presenti rispettivamente alle porte A0 e A2. Una volta registrate in un file al termine dell'acquisizione, esse sono trattate secondo Eq. 1 in modo da poter ricostruire per punti la curva caratteristica di collettore alla data corrente di base. La misura può essere poi ripetuta per altri valori di I_B per ottenere una famiglia di curve.

Affinché l'operazione abbia l'esito progettato, occorre rispettare un certo numero di requisiti nella forma d'onda prodotta dal generatore di funzioni. Questi requisiti, elencati qui di seguito, vanno controllati accuratamente *prima* di avviare l'acquisizione, usando l'oscilloscopio e, preferibilmente, avendo scollegato gli ingressi di Arduino.

1. Arduino può leggere e digitalizzare solo tensioni positive (o nulle) rispetto alla linea di terra. Di default il generatore di funzioni produce forme d'onda alternate, che cioè corrispondono a d.d.p. positive e negative. Pertanto è necessario aggiungere un offset continuo al segnale alternato. Questo si ottiene facilmente agendo sulla manopola **OFFSET** del generatore di funzioni, che deve essere estratta e regolata per aggiungere l'offset desiderato.
2. Arduino, poi, digitalizza segnali che valgano al massimo la propria V_{ref} , cioè, nelle nostre condizioni di operazione, circa 5 V. Pertanto è necessario che l'ampiezza massima della forma d'onda triangolare prodotta sia circa 5 V, o un po' meno per sicurezza. Questo si ottiene lavorando con la manopola **AMPL** del generatore di funzioni, iterando, se necessario, la regolazione dell'offset. Alla fine, sull'oscilloscopio usato per monitorare la forma d'onda essa deve apparire come un triangolo con ampiezza compresa tra circa 0 e circa 5 V.
3. La frequenza dell'onda triangolare deve anche essere aggiustata in maniera opportuna. Come chiariremo nel seguito, al massimo il campione acquisito consta di 256 punti. L'intervallo di campionamento Δt , nella configurazione usata per l'esperienza, può essere variato in passi discreti tra 1 e 9 ms (al solito, la selezione si esegue via software nello script di Python che controlla l'operazione), per cui l'intera acquisizione, trascurando gli eventuali ritardi e latenze di Arduino, dura al massimo un tempo compreso tra circa 250 ms e circa 2.3 s. Questo tempo deve corrispondere a quello necessario perché la forma d'onda triangolare passi dal suo valore minimo a quello massimo, cioè a metà del periodo T dell'onda stessa. Si sottolinea che le frequenze necessarie per la corretta esecuzione dell'esperienza sono così basse da rendere difficoltosa la visualizzazione di tracce continue sullo schermo dell'oscilloscopio, per cui si consiglia vivamente di lasciare la regolazione della frequenza “per ultima”, in modo da eseguire più agevolmente il monitoraggio dell'ampiezza.

Soddisfare questi requisiti, in maniera più o meno rigorosa a seconda delle proprie capacità e della propria pazienza, è condizione essenziale per eseguire la misura, ma non sufficiente. Infatti occorre anche fare in modo che l'acquisizione dei dati da parte di Arduino parta *in sincrono* con l'"inizio" della forma d'onda triangolare. In altre parole, Arduino deve essere "triggerato" in modo opportuno dal generatore di funzioni.

C. Trigger

Di come sia possibile triggerare Arduino (per i nostri scopi) ci siamo già occupati in precedenza. Ripetiamo qui alcune considerazioni generali e le ricette pratiche impiegate, partendo dalla constatazione che, in linea di principio, esistono vari modi per eseguire la sincronizzazione richiesta. Per esempio, potremmo fare come si fa nei circuiti di trigger degli oscilloscopi, cioè monitorare continuamente il segnale V_1 e far partire l'acquisizione solo se questo è nella sua fase crescente. L'implementazione pratica di questo trigger è però abbastanza poco affidabile: essa è relativamente lenta (occorrono almeno due campionamenti consecutivi per capire la pendenza della forma d'onda) e, soprattutto, sensibile a fluttuazioni di lettura dovute a rumore, che potrebbero facilmente ingannare il trigger. La strategia di sincronizzazione qui impiegata si basa invece sulla disponibilità di un segnale specifico, progettato proprio per scopi di questo tipo e prodotto dal generatore di funzioni in sincrono con le forme d'onda. Questo segnale si trova su un connettore BNC marcato TTL/CMOS OUTPUT sul retro della scatola dell'apparecchio o, per alcuni modelli, direttamente sul frontale. L'indicazione TTL [4] significa che questo segnale segue le prescrizioni dello standard TTL [5] in termini di livelli. Questo standard stabilisce che il livello è basso ("low", corrispondente a uno 0 binario) in corrispondenza a una d.d.p. praticamente nulla (ovvero minore di 0.8 V), e che il livello è alto ("high", corrispondente a un 1 binario) per una d.d.p. maggiore di 2 V. Nella pratica (e quasi sempre), livello basso e livello alto di un segnale TTL corrispondono rispettivamente a circa 0 e circa 5 V (informazione molto utile, da ricordare sempre).

Questo segnale può essere facilmente letto da Arduino in forma *digitale*, cioè usando come ingresso una delle sue porte digitali (nella pratica è la porta 5). L'operazione è vantaggiosa rispetto alla lettura di un segnale analogico, poiché i livelli sono codificati, e pertanto meno sensibili a rumore e fluttuazioni, e non c'è bisogno di introdurre delle istruzioni software di confronto per stabilire se ci si trova al livello basso o alto [6].

Come già accennato, questo segnale TTL è sincrono con la forma d'onda [7]: il timing è infatti quello rappresentato in Fig. 2, dove si osserva come il segnale presente all'uscita del generatore di funzioni, opportunamente regolato secondo i requisiti prima elencati, inizi la sua fase di salita quando il segnale TTL passa dal livello alto a

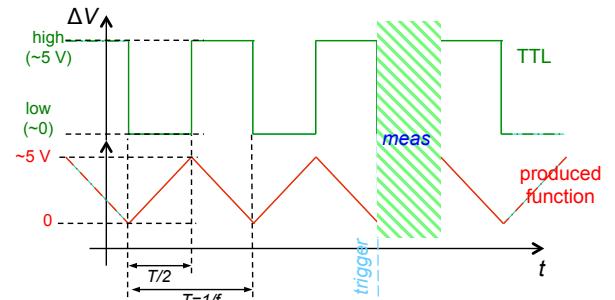


Figura 2. Schema del timing dei segnali: la forma d'onda triangolare prodotta dal generatore e usata come V_1 nell'esperienza è rappresentata in rosso, il segnale TTL di sincronizzazione in verde. Nella figura si suppone che l'istante di trigger sia quello indicato e che la misura avvenga all'interno del riquadro ombreggiato.

quello basso, cioè da circa 5 V a circa 0. Allora nel loop di acquisizione scritto nello sketch di Arduino dovremo fare in modo che le misure abbiano inizio quando il livello letto dalla porta digitale 5 passa da alto a basso. Notate, infatti, che porre una semplice condizione sul livello della porta digitale stessa, stabilendo per intenderci che l'acquisizione inizi quando esso è basso, potrebbe far partire l'acquisizione in un *qualsiasi* istante in cui il segnale TTL è basso, cioè non necessariamente quando esso passa da alto a basso, come invece è richiesto qui.

La Fig. 3 mostra lo schema delle connessioni che devono essere fatte alla scheda di Arduino. In sostanza è necessario impiegare quattro boccole a banana di colore diverso:

1. la boccola nera deve essere collegata alla linea di terra, o massa, del circuito di Fig. 1(b);
2. la verde deve ricevere il segnale di sincronismo proveniente dall'uscita (connettore BNC) TTL/CMOS OUTPUT sul retro, o sul frontale, del generatore di funzioni;
3. la gialla deve essere collegata al collettore del transistor [C in Fig. 1(b)] allo scopo di leggere la d.d.p. V_2 ;
4. la blu deve essere collegata a valle del resistore R_C di Fig. 1(b) allo scopo di leggere V_1 e anche all'uscita del generatore di funzioni che fornisce V_1 .

La boccola volante rossa è non collegata: tuttavia essa può essere utile, dato che corrisponde alla porta digitale 7 di Arduino, configurata, per default, come un'uscita. Un'opportuna istruzione dello sketch la pone costantemente a livello alto: dunque la misura della d.d.p. presente su questa porta (rispetto alla linea di massa, o terra) consente di determinare il valore della tensione di riferimento V_{ref} usata da Arduino, che può essere utile qualora si voglia stabilire il fattore di calibrazione ξ che serve per convertire la lettura digitalizzata in unità di differenza

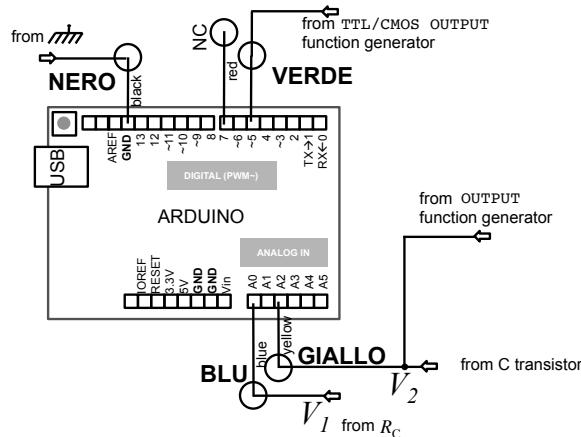


Figura 3. Schema delle connessioni da effettuare alla scheda Arduino con indicazione del colore delle boccole volanti. L’eventuale lettura di V_{ref} , utile per la calibrazione, può essere effettuata eseguendo una misura di tensione tra porta 7 (boccola rossa) e massa; infatti la porta 7 viene posta a livello alto, in prima approssimazione corrispondente a V_{ref} , da un’opportuna istruzione nello sketch.

di potenziale. In questa esperienza non è strettamente necessario individuare le grandezze richieste (V_{CE} e I_C) con grande accuratezza, per cui l'approccio di calibrazione "alternativo" basato sulla misura di V_{ref} (cioè della d.d.p. presente tra porta digitale 7 e linea di massa) è ampiamente sufficiente. Addirittura, vi mancasse il tempo, per evitare di prolungare troppo la vostra permanenza in laboratorio potreste anche accontentarvi di una sorta di calibrazione nominale, ponendo $\xi = 5 \text{ V} / 1023 \approx 5 \text{ mV}$.

D. Numero di punti acquisiti

Un altro aspetto di interesse per la gestione dell'esperienza riguarda il numero di punti sperimentali registrati in ogni acquisizione, cioè per ogni regolazione della corrente di base I_B . Come abbiamo già avuto modo di discutere in altre occasioni, la ridotta disponibilità di memoria del microcontroller di Arduino impone dei limiti nel massimo numero di dati che possono essere registrati, che qui è fissato a 256 punti [cioè vengono registrati su file due colonne con al massimo 256 punti ognuna, corrispondenti alla lettura digitalizzata rispettivamente delle porte analogiche A2 e A0, ovvero dei segnali V_2 e V_1 di Fig. 1(b)]. Il nostro desiderio sarebbe ovviamente che questi 256 punti fossero relativi all'*intera* fase di salita dell'onda triangolare usata come V_1 , in modo da poter esplorare l'intero intervallo di d.d.p. a nostra disposizione. Ottenere o meno questo risultato dipende dalla scelta dell'intervallo di campionamento Δt e della frequenza f del generatore di funzioni.

Infatti l'intervallo di campionamento Δt determina la durata temporale complessiva (massima) dell'acquisizione, $T_{acq} \simeq 256 \times \Delta t$ (diventa un'uguaglianza se si trascu-

rano latenze, ritardi e fluttuazioni negli intervalli temporali), che deve essere confrontata con il semiperiodo $T/2 = 1/(2f)$ della forma d'onda prodotta dal generatore di funzioni. Possono allora verificarsi queste situazioni:

1. per T_{acq} (nettamente) minore di $T/2$, la misura viene effettuata in corrispondenza a un intervallo di variazione della d.d.p. V_1 (nettamente) minore del massimo possibile, cioè, in altre parole, V_1 è limitata in alto a qualcosa di (nettamente) minore del massimo possibile (circa 5 V);
 2. viceversa, se T_{acq} è (nettamente) maggiore di $T/2$, le misure vengono eseguite non solo per la fase in cui V_1 cresce, ma anche per quella in cui esso cala; di conseguenza i dati registrati non sono più ordinati in maniera monotona. Per evitare gli eventuali problemi connessi a questa situazione [8], l'acquisizione di Arduino viene interrotta automaticamente non appena il livello del segnale TTL di sincronismo passa per la prima volta dal livello alto a quello basso, come rappresentato in Fig. 2: dunque in ogni caso i dati vengono registrati solo nella fase crescente di V_1 .

L'ultima affermazione che abbiamo riportato implica che, nel caso considerato, il campione di misure sia costituito da un numero minore di 256 punti. Quindi rimane importante che l'intervallo di campionamento e la frequenza del generatore di funzioni vengano scelte in modo opportuno. Come illustreremo nel seguito, per aiutare nella verifica lo script di Python che gestisce l'acquisizione riporta sulla console delle indicazioni utili, cioè il numero di punti acquisito, che idealmente dovrebbe essere prossimo a, 256, e anche i valori minimi e massimi digitalizzati per V_1 , che dovrebbero essere prossimi rispettivamente a 0 e 1023.

III. SKETCH E SCRIPT

Lo sketch di Arduino scritto per l'esperienza, disponibile in rete sotto il nome `curv.ino` e riportato in Appendice, è sviluppato sulla base di quello, discusso in precedenza, usato per la curva caratteristica del diodo a giunzione. Le modifiche principali riguardano:

1. la definizione della porta 5 come ingresso digitale (l'istruzione, posta nella parte di inizializzazione della scheda, recita `pinMode(sincPin, INPUT);`, dove `sincPin` è una costante intera che punta alla porta in questione);
 2. l'implementazione del trigger come descritto sopra, che in sostanza è eseguito attraverso due loop successivi di attesa realizzati con le istruzioni `while (sinc==LOW)` e `while (sinc==HIGH)`, dove la lettura della porta di ingresso digitale viene eseguita con le istruzioni `sinc = digitalRead(sincPin);` inserite nei loop;

3. la presenza dell'istruzione `if (sinc==HIGH) break;` all'interno del ciclo di acquisizione, che serve per interromperlo se il segnale di sincronismo passa a livello alto, secondo quanto discusso prima;
4. l'eventuale aggiunta della parola `'9999'` al termine dell'array che contiene i dati nel caso in cui l'acquisizione sia eseguita su meno di 256 punti, che serve per segnalare allo script di Python la fine dei dati.

Anche lo script di Python, disponibile in rete sotto il nome `curv_v1.py`, è molto simile a quello già impiegato per l'esperienza della curva caratteristica del diodo. Si ricorda che, al termine dell'acquisizione, viene creato un file di due colonne (e un numero di linee pari a quello delle misure effettuate, fino a un massimo di 256) che contiene i valori digitalizzati di, nell'ordine, V_1 e V_2 [vedi Fig. 1(b)]. Questi dati sono disponibili per ulteriori analisi e in particolare per tracciare per punti la curva caratteristica di interesse. Inoltre lo script, oltre al nome del file e all'eventuale directory di archiviazione, permette di impostare l'intervallo di campionamento Δt in unità di ms (da 1 a 9 ms, con passi discreti di 1 ms). Un'opportuna serie di istruzioni consente di gestire e registrare files contenenti meno di 256 punti: a questo scopo i dati trasmessi da Arduino via porta seriale vengono monitorati via via che arrivano, e il loop di lettura e registrazione su file di questi dati viene interrotto o quando sono stati registrati 256 coppie di dati, oppure quando il dato è rappresentato dalla parola `'9999'`. Il numero di punti acquisiti, assieme al valore minimo e massimo digitalizzato di V_1 , viene scritto sulla console per praticità.

IV. ESEMPIO DI MISURE

La Fig. 4 riporta come esempio una famiglia di curve caratteristiche di uscita, o di collettore, per il transistor 2N1711, registrate come descritto in questa nota per le intensità di corrente di base I_B riportate in legenda (nell'intervallo 1.2–35 μ A, che è sufficiente per i nostri scopi). Le acquisizioni sono state effettuate con $\Delta t = 1$ ms (nominale); i campioni sono costituiti da 218 punti (dunque un numero abbastanza prossimo, ma non troppo, a quello “ideale” di 256), con V_1 variabile tipicamente tra 1 e 988 digit (dunque un intervallo abbastanza prossimo a quello “ideale” compreso tra 0 e 1023).

Per la conversione delle letture digitalizzate in unità di d.d.p. (V) si è usato il fattore di conversione, $\xi = (4.98 \pm 0.03)$ mV/digit, stabilito leggendo la V_{ref} di Arduino presente sulla porta 7 e tenendo conto che il convertitore analogico/digitale di Arduino ha una “dinamica” di 10 bit. Per la determinazione dell'intensità di corrente I_C secondo Eq. 1 si è impiegato il valore, misurato con tester digitale, $R_C = (1028 \pm 8)$ ohm. Le barre di errore del grafico su sono state deliberatamente e arbitrariamente scelte come rappresentative della sola incertezza di digitalizzazione (± 1 digit, convertito in unità fisiche per la misura di V_{CE} , per l'asse orizzontale e

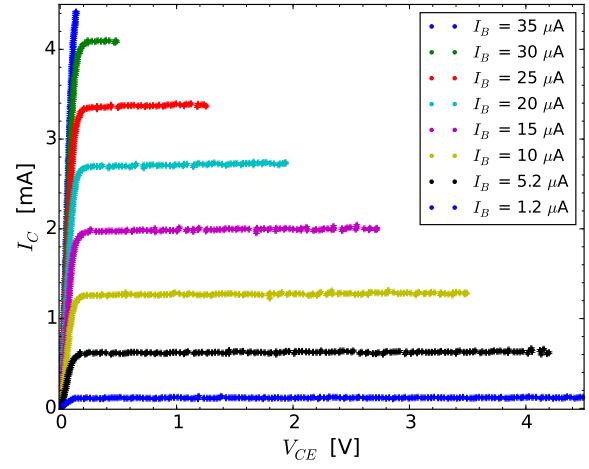


Figura 4. Esempio di famiglia di curve caratteristiche di uscita del transistor 2N1711 ricostruite nell'esperienza pratica. Le barre di errore tengono solo conto dell'incertezza di digitalizzazione; l'incertezza di calibrazione è attesa contribuire per meno dell'1% della lettura per le barre di errore orizzontali, e per meno del 2% della lettura per le barre di errore verticali.

± 2 digit, convertito in unità fisiche per la misura di I_C , per l'asse verticale, dove si è tenuto conto che la misura di corrente è proporzionale alla differenza di due distinte letture di d.d.p. e si è usata l'incertezza massima risultante). Questa scelta è sostanzialmente dovuta al fatto che i dati acquisiti non vengono qui impiegati per trarre conclusioni quantitative, per esempio per determinare una grandezza fisica tramite best-fit, per cui non si ritiene utile tenere conto degli errori di calibrazione. La scelta è debitamente segnalata nella caption della figura, dove si riporta anche una stima dell'errore di calibrazione.

Le curve ottenute sono in accordo qualitativo con le aspettative. In particolare, per $I_B = 1.2 \mu$ A la corrente che fluisce nel collettore è al massimo dell'ordine di poche decine di μ A, suggerendo come in queste condizioni il transistor si trovi in condizioni di lavoro prossime a quelle di *interdizione*. Infatti la debole corrente di base misurata corrisponde a V_{BE} inferiore, o perlomeno non decisamente superiore, alla soglia. All'aumentare di I_B , invece, la giunzione BE diventa sicuramente polarizzata direttamente. In queste condizioni, si verifica che il transistor può accedere al regime attivo, in cui la giunzione BC è polarizzata inversamente, cioè, con ovvio significato dei simboli, $V_{BC} < V_{thr}$. Poiché, con altrettanto ovvio significato dei simboli, $V_{CE} = V_{BE} + V_{CB} = V_{BE} - V_{BC}$, si può porre $V_{BC} = V_{BE} - V_{CE}$, per cui la giunzione BC risulta polarizzata inversamente per $V_{CE} > V_{BE} - V_{thr}$. D'altra parte, la polarizzazione diretta della giunzione BE implica che V_{BE} sia “leggermente” al di sopra di V_{thr} , per cui la condizione che stiamo cercando si traduce in $V_{CE} > V'$, tipicamente con $0 < V' < V_{thr}$ (e V' ovviamente dipendente da V_{BE} , cioè da I_B). La Fig. 4 mostra che per il transistor esaminato, e per il range di I_B esplorato,

rato, si ha $V' \sim 0.2$ V; infatti al di sotto di questo valore I_C tende rapidamente a zero, al di sopra di questo valore I_C diventa sensibilmente maggiore di zero, mostrando un'evidente transizione dal regime di *saturazione* a quello attivo.

È rilevante osservare che le curve acquisite con i valori più alti di I_B esplorati in questa esperienza non coprono lo stesso range di variazione di V_{CE} di quelle registrate a valori di I_B minori. La spiegazione è ovvia e immediata: quando la giunzione BE è “fortemente” polarizzata in modo diretto, allora una rilevante intensità di corrente scorre attraverso il collettore. Questa corrente provoca una altrettanto rilevante caduta di potenziale sulla resistenza di collettore R_C , e quindi limita il massimo valore possibile per V_{CE} . In altre parole, il generatore che produce V_{CE} , visto in termini di Thévenin, ha una resistenza interna elevata, che è data dalla somma della resistenza interna del generatore di funzioni (50 ohm, quindi trascurabile) e di R_C (vale circa 1 kohm, tutt’altro che trascurabile).

Per il transistor esaminato, a $I_B = 35 \mu\text{A}$ la corrente I_C diventa ben presto così alta da rendere la caduta di potenziale su R_C simile a V_1 , per cui, di fatto, può essere ricostruita solo una piccola porzione della curva caratteristica. Infatti, tenendo conto che $R_C = 1$ kohm nominali, per $I_C \rightarrow 5 \text{ mA}$, $V_{CE} \rightarrow 0$ anche per $V_1 \sim 5 \text{ V}$. La limitazione potrebbe essere rimossa impiegando una resistenza di collettore più bassa, ma questo non è necessario per i nostri scopi.

Il brusco passaggio da piccole a grandi correnti di collettore può ovviamente essere realizzato anche mantenendo costante V_1 e agendo su I_B : nelle condizioni dell’esperienza, per $V_1 \sim 5 \text{ V}$, per esempio, quando I_B viene variata da circa $1 \mu\text{A}$ a $35 \mu\text{A}$ la corrente di collettore passa da un valore trascurabile (poche decine di μA) a valori dell’ordine di 5 mA . Questo comportamento è alla base dell’impiego del transistor come *switch* e illustra bene come la variazione di una “piccola” corrente (quella di base, e la variazione è piccola in termini assoluti) possa controllare la variazione di una corrente ben più “grande” (quella di collettore, e la variazione è grande in termini assoluti).

Il modello di Ebers-Moll, che, ricordiamo, rappresenta una descrizione approssimativa del funzionamento del transistor, prevede

$$I_C = -I_{0C} \left[\exp \left(\frac{V_{BC}}{\eta V_T} \right) - 1 \right] + \alpha_F I_{0E} \left[\exp \left(\frac{V_{BE}}{\eta V_T} \right) - 1 \right], \quad (3)$$

dove I_{0C} e I_{0E} sono le correnti di saturazione inversa delle due giunzioni BC e BE, e $\alpha_F \simeq 1$ è il coefficiente dell’“effetto transistor”. Nel regime di cui ci stiamo occupando, in cui BE è polarizzata direttamente, il secondo termine della somma fornisce un contributo pressoché costante e numericamente molto rilevante, essendo $V_{BE} \gg \eta V_T$. Il primo termine della somma diventa sicuramente trascurabile quando V_{BC} è piccolo, o addirittura negativo, come si verifica per $V_{CE} > V_{BE}$, cioè in pieno regime attivo. Nel regime di saturazione, invece, V_{BC} diventa positiva e approssima V_{thr} per $V_{CE} \rightarrow 0$: in

queste condizioni il primo termine della somma in Eq. 3 contribuisce con un termine negativo, che diventa tanto più grande quanto più V_{BC} si avvicina a V_{thr} . Di conseguenza I_C tende ad annullarsi seguendo un brusco andamento, come trovato in Fig. 4 per $V_{CE} < V'$. In altre parole, l’Eq. 3 può giustificare analiticamente l’andamento delle curve caratteristiche nella transizione da regime attivo a regime di saturazione, ma il suo impiego, per esempio attraverso un best-fit, *non è richiesto* nell’esperienza pratica. Esso, infatti, non è particolarmente significativo, poiché i parametri di fit eventualmente determinati, che sono legati in maniera non completamente ovvia con i parametri fisici del modello, possono essere trovati in maniera generalmente più immediata [9].

A. Guadagno in corrente

Una considerazione immediata che si può fare esaminando le curve caratteristiche di uscita di un transistor è la differenza di ordini di grandezza tra le intensità di corrente di base e di collettore che si verifica quando si opera in regime attivo. Questa è in effetti una delle caratteristiche più importanti, o la più importante, del transistor a giunzione bipolare (*BJT*). Infatti l’effetto transistor, specifico del regime attivo, dà luogo a un’amplificazione delle correnti (continue) secondo la legge $I_C = \beta_F I_B$. Il termine β_F , definito spesso anche h_{FE} , è legato a quello del coefficiente α_F dell’effetto transistor secondo la $\beta_F = \alpha_F / (1 - \alpha_F)$; poiché $\alpha_F \sim 1$, β_F è tipicamente grande, dell’ordine delle decine o centinaia.

In prima approssimazione, β_F può essere ritenuto indipendente da I_C . Tuttavia, per grandi correnti di collettore l’effetto transistor diventa più marcato, poiché la probabilità che gli elettroni che superano l’interfaccia emettitore-base si ricombinino nella base diminuisce a causa della densità finita di lacune, conseguenza del drogaggio a bassa densità di questa regione (drogata di tipo *p* in un transistor *npn*). Pertanto, α_F si avvicina sempre di più all’unità, e β_F tende ad aumentare con I_C . Sfortunatamente, poi, il valore di α_F , oltre a dipendere dalle condizioni di operazione (temperatura inclusa), ha un’estrema variabilità legata alle fluttuazioni di fabbricazione. Per intenderci, il datasheet del transistor 2N1711 indica, per $I_C = 10 \text{ mA}$ e $V_{CE} = 10 \text{ V}$ (condizioni non raggiunte nella presente esperienza), un valore *tipico* $\beta_F = 80$, ma, allo stesso tempo, avverte che la tolleranza di fabbricazione rende possibili valori compresi tra $\beta_F = 35$ e $\beta_F = 300$ (uno spread di quasi un ordine di grandezza).

La misura della curva caratteristica di uscita permette di determinare β_F nelle varie condizioni di operazione. Allo scopo, è sufficiente prendere i dati relativi a un certo valore di I_B (misurato con il multmetro), scegliere un dato valore di V_{CE} (la scelta è “libera”, qui si è preso $V_{CE} = 1.0 \text{ V}$ nominali), vedere nel file o, per praticità, direttamente dal grafico, quanto vale I_C e infine calcolare il rapporto $\beta_F = I_C / I_B$. Come esempio, si riporta in

I_B [μ A]	I_C [mA]	β_F
5.2 ± 0.1	0.63 ± 0.02	121 ± 5
10.1 ± 0.2	1.32 ± 0.04	131 ± 8
14.9 ± 0.2	1.99 ± 0.06	133 ± 10
19.8 ± 0.3	2.75 ± 0.08	139 ± 12

Tabella I. Tabella I. Valori di I_B e I_C desunti da alcune delle curve caratteristiche di Fig. 4 e corrispondente valore del guadagno in corrente (continua) $\beta_F = I_B/I_C$. I dati si riferiscono tutti al valore $V_{CE} = 1.0$ V nominali. Le incertezze sulla misura di I_B sono quelle dovute all'impiego del multimetro digitale, quelle su I_C tengono conto dell'errore di digitalizzazione, di quello di calibrazione di Arduino e dell'errore nella misura di R_C .

Tab. I il calcolo eseguito per alcuni valori, naturalmente tutti corrispondenti all'operazione nel regime attivo (il guadagno in corrente non è definito per gli altri regimi di operazione del transistor). Per la determinazione dell'incertezza su I_C , e quindi su β_F , sono state considerate tutte le cause di errore (calibrazione di V_1 e V_2 , loro incertezza di digitalizzazione, incertezza sulla misura di R_C , il tutto opportunamente propagato con somme in quadratura). Si osserva come i valori di β_F tendano a crescere, appena più delle barre di errore, con I_C .

Come è noto, il comportamento da amplificatore in corrente del transistor in regime attivo può essere leggermente differente a seconda che le correnti di base e collettore siano continue o alternate (in quest'ultimo caso, ci si riferisce a *deboli segnali alternati*). Infatti il guadagno in corrente corrispondente, indicato con β_f (o h_{fe} , dove l'uso dei pedici minuscoli rimanda alla presenza dei "deboli" segnali), è riportato nel datasheet avere un valore tipico di 135 (tolleranza da 70 a 300, dunque sempre gigantesca). La sua misura diretta richiede di realizzare un opportuno circuito amplificatore e l'uso di un debito modello, secondo quanto sarà svolto in una prossima esperienza pratica. Qui possiamo limitarci a darne una stima. A questo scopo possiamo registrare curve caratteristiche I_C vs V_{CE} in corrispondenza a valori fissati di I_B che siano "poco distanti" l'uno dall'altro. In questo esempio, sono state impiegate le curve corrispondenti a $I_B = (10.1 \pm 0.2)$ μ A e $I_B = (14.9 \pm 0.2)$ μ A [10], di cui sono stati ricavati i valori di I_C per $V_{CE} = 1.0$ V (nominali) riportati in Tab. I. Quindi si è definito $\beta_f \simeq \Delta I_C / \Delta I_B$, dove ΔI_B e ΔI_C sono le variazioni delle due intensità di corrente corrispondenti alle due condizioni (esse si determinano facilmente eseguendo la differenza). Si è ottenuto $\beta_f = 140 \pm 12$, un valore leggermente superiore rispetto ai corrispondenti β_F . Dunque il transistor esaminato, debitamente polarizzato, si comporta da amplificatore con un guadagno grossolanamente compreso tra 130 e 140 sia per correnti continue che per (deboli) correnti alternate.

B. Effetto Early

Una delle affermazioni approssimate che si usa fare per descrivere il comportamento di un transistor nel regime attivo è che la corrente I_C dipende *esclusivamente* da I_B . Questo è anche il comportamento grossolanamente previsto dal modello di Ebers-Moll: se nell'Eq. 3 poniamo condizioni tipiche del regime attivo, cioè $V_{BE} \gtrsim V_{thr}$ e $V_{BC} < 0$, I_C risulta pressoché indipendente da V_{CE} (la dipendenza da V_{CE} è implicita, aumentando V_{CE} si rende sempre più negativa V_{BC} , ma poiché V_{BC} compare come argomento di un esponenziale, la variazione del termine che la contiene è poco rilevante).

Che questa approssimazione sia non completamente verificata si può dimostrare con i nostri dati. Infatti risulta abbastanza evidente che, anche in pieno regime attivo, I_C non è costante, ma tende a crescere con V_{CE} . L'effetto che interpreta questo andamento è chiamato *effetto Early*. All'aumentare di V_{CE} , la giunzione base-collettore risulta inversamente polarizzata in modo sempre più marcato e il campo elettrico nella regione di collettore (orientato "verso la base" per un transistor *npn*) aumenta la sua intensità. Dal punto di vista qualitativo, possiamo supporre che:

- la regione di svuotamento della giunzione BC aumenti la sua estensione longitudinale, ovvero il suo spessore, rendendo meno probabile la ricombinazione della corrente di elettroni che proviene dall'emettitore e passa nella base;
- l'efficacia della raccolta di elettroni da parte del campo presente nel collettore aumenti, e una densità maggiore di carica, cioè una corrente di intensità maggiore, fluisca nel collettore.

Entrambi questi effetti portano α_F più prossimo all'unità, cioè aumentano β_F e di conseguenza l'intensità di corrente I_C , al crescere di V_{CE} .

L'effetto Early è modellato dalla seguente espressione:

$$I_C \simeq \beta_F I_B \left(1 + \frac{V_{CE}}{V_{Early}} \right), \quad (4)$$

dove V_{Early} , che ha le dimensioni di una d.d.p., rappresenta un parametro caratteristico del transistor impiegato (naturalmente dipende anche dalle condizioni di operazione) e ha valori tipici compresi tra circa 50 e circa 200 V. Supponendo costante il termine $\beta_F I_B$, l'Eq. 4 stabilisce un andamento lineare crescente, dove $-V_{Early}$ rappresenta l'intercetta con l'asse orizzontale del grafico della curva caratteristica di uscita.

La Fig. 5 mostra gli stessi dati di Fig. 4 per $I_B = 10$ μ A, graficati su una diversa scala per evidenziare l'andamento crescente. I dati per $V_{CE} > 0.2$ V (scelto arbitrariamente) sono stati fittati numericamente a una retta del tipo $y = a + bx$; supponendo (arbitrariamente) che le incertezze ΔI_C fossero correttamente rappresentative dell'errore statistico (esse corrispondono all'errore

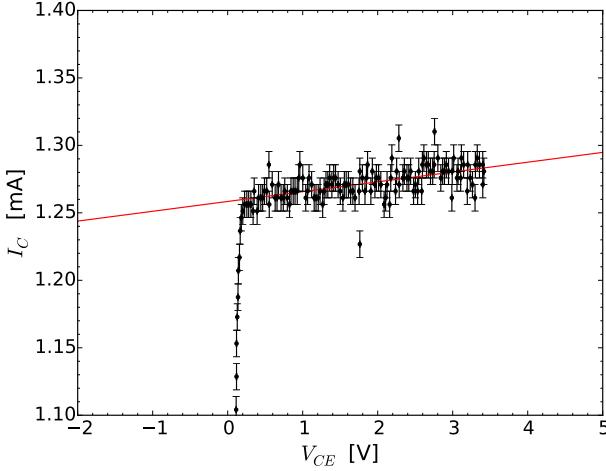


Figura 5. Stessi dati di Fig. 4 per $I_B = 10 \mu\text{A}$ graficati su una scala diversa per evidenziare l'andamento crescente attribuito all'effetto Early. La linea continua rappresenta un best-fit lineare ai dati corrispondenti a $V_{CE} > 0.2$ V. Si notano alcuni punti sperimentali apparentemente fuori dalle aspettative, che potrebbero essere dovuti a fluttuazioni nella digitalizzazione o contributo di rumori esterni.

di digitalizzazione), nella procedura di best-fit è stata scelta l'opzione `absolute_sigma=True`. Il fit ottenuto è sovrapposto ai dati in Fig 5. I risultati sono

$$a = (1.258 \pm 0.002) \text{ mA} \quad (5)$$

$$b = (7.3 \pm 9) \mu\text{A/V} \quad (6)$$

$$\chi^2/\text{ndof} = 114/120 \quad (7)$$

$$R^2 = 0.999 \quad (8)$$

$$\text{norm. cov.} = -0.89 . \quad (9)$$

I parametri trovati conducono a $V_{Early} = a/b = (173 \pm 22)$ V, dove nell'incertezza abbiamo tenuto conto della calibrazione di Arduino e della barra di errore nella misura di R_C .

APPENDICE: SKETCH

```
//Dichiarazione
const unsigned int sincPin = 5; //pin 5 ingresso digitale per la sincronizzazione con il generatore
const unsigned int digitalPin = 7; //pin 7 uscita digitale (serve per eventuale calibrazione convertitore)
const unsigned int analogPin_0=0; //pin A0 per lettura V1
const unsigned int analogPin_2=2; //pin A2 per lettura V2
unsigned int i=0; //variabile di conteggio per i cicli
int V1[256]; //array per memorizzare V1 (d.d.p, letta da analogPin_0)
int V2[256]; //array per memorizzare V2 (d.d.p, letta da analogPin_2)
int delayms; //variabile che contiene il ritardo tra due step successivi (in unita' di 1 ms, fino a 9 ms)
int start=0; //flag per dare inizio alla misura
int sinc; //variabile di sincronizzazione
int nmis=0; //variabile con il numero di punti acquisiti
//Inizializzazione
void setup()
{
    pinMode(sincPin, INPUT); //pin sincPin configurato come ingresso digitale
    Serial.begin(9600); //inizializzazione della porta seriale
    Serial.flush(); // svuota il buffer della porta seriale
}
//Ciclo di istruzioni del programma
void loop()
{
    if (Serial.available() >0) // Controlla se il buffer seriale ha qualcosa
    {
        delayms = (Serial.read()-'0')*1; // Legge il byte e lo interpreta come ritardo (unita' 1 ms)
        Serial.flush(); // Svuota il buffer della seriale
    start=1; // Pone il flag start a uno
    }
    if(!start) return // solo se il flag e' a uno parte l'acquisizione
```

```

delay(2000); // attende 2s per evitare casini
digitalWrite(digitalPin,HIGH); // pone digitalPin a livello alto (per eventuale calibrazione)
sinc = digitalRead(sincPin);//legge sincPin
while (sinc==LOW) // ciclo di attesa iniziale per sincronizzazione, attende che sincPin vada alto
{sinc = digitalRead(sincPin);} //legge sincPin
while (sinc==HIGH) // ciclo di attesa iniziale per sincronizzazione, attende che sincPin vada basso
{sinc = digitalRead(sincPin);} //legge sincPin
for(i=0;i<256;i++) //ciclo di acquisizione temporizzata delle misure (fino a 256 punti)
{
    sinc = digitalRead(sincPin); //legge sincPin
    if (sinc==HIGH) break; //esce dal ciclo se sincPin torna alto
    delay(delayms); //aspetta il tempo impostato
    V1[i]=analogRead(analogPin_uno); //legge il pin analogPin_uno
    V2[i]=analogRead(analogPin_due); //legge il pin analogPin_due
}
nmis = i; //scrive in nmis il numero effettivo di punti acquisiti
for(i=0;i<nmis;i++) //nuovo ciclo che scorre gli array di dati e li scrive sulla seriale
{
    Serial.print(V1[i]);
    Serial.print(" ");
    Serial.println(V2[i]);
}
if (nmis<256) //se il numero di punti e' minore di 256 scrive un carattere convenzionale di fine file
{
    Serial.println(9999); //il carattere convenzionale e' 9999
}
start=0; // Annulla il flag
Serial.flush(); // svuota il buffer della porta seriale
}

```

- [1] La tensione di soglia per un'ordinaria giunzione in silicio vale normalmente $V_{thr} = 0.45 - 0.65$ V. Notate che il suo valore non è completamente definito: da un lato, le tolleranze costruttive e i parametri di disegno possono influenzare V_{thr} , dall'altro, come abbiamo già osservato per il diodo, la grandezza V_{thr} non è individuabile in modo univoco. Dunque considerate sempre come approssimativo il valore di V_{thr} ogni volta che ci si fa riferimento.
- [2] A differenza di V_{BE} , non ci sono limiti di principio nel range di variazione di V_{CE} . Infatti è sufficiente disporre di un generatore di d.d.p. positiva che vari da zero (convenzionalmente le curve caratteristiche partono da $V_{CE} = 0$) fino a un valore massimo che è idealmente limitato dalla massima tensione, o corrente, sopportabile dal transistor. La disponibilità di strumentazione che abbiamo in laboratorio è sicuramente adeguata per evitare di lavorare in condizioni che possano danneggiare il transistor 2N1711.
- [3] Sapete per esperienza che il potenziometro è un dispositivo "delicato": la rotazione dell'alberino può comportare brusche variazioni della d.d.p. in uscita dal partitore, e questa d.d.p. può fluttuare anche in maniera sensibile. È compito del "bravo sperimentale" evitare che queste difficoltà tecniche diventino troppo importanti e anche che esse siano causa di eccessive (e ovvie) lamentazioni.
- [4] In realtà il segnale può anche seguire lo standard cosid-

detto CMOS, che è caratterizzato da livelli ed eventualmente polarità diverse. Per commutare da TTL a CMOS occorre agire su una specifica manopola del frontale del generatore di funzioni, estraendola e quindi ruotandola per aggiustare i livelli.

- [5] TTL sta per *Transistor-Transistor-Logic*, uno standard di impulsi logici tra i primi a essere codificato, già agli albori dell'elettronica digitale.
- [6] La lettura di segnali "digitali" è quasi sempre preferibile rispetto a quella di segnali analogici nel caso in cui si sia interessati solo a conoscere uno "stato" (alto o basso, in questo caso). Tuttavia vale la pena di ricordare che possono esserci delle eccezioni nel caso di segnali che passano da un livello all'altro in modo molto rapido o che hanno una durata molto breve: essi potrebbero non essere letti in modo corretto da Arduino o dare luogo a "rimbalzi" del segnale. Al momento, questi problemi non sembrano essere presenti nell'implementazione pratica qui proposta, grazie anche alla relativa lentezza del campionamento, ma non se ne può escludere a priori la presenza.
- [7] L'eventuale ritardo interno al generatore di funzioni tra le forme d'onda di interesse è sicuramente trascurabile per gli scopi di questa esperienza. I ritardi interni delle apparecchiature sperimentali potranno diventare rilevanti quando eseguirete misure su scale temporali brevi, per

esempio inferiori al μs (cosa che non si realizza nel nostro corso di laboratorio).

- [8] Quello qui menzionato è un requisito di carattere poco più che “estetico”. Sarebbe infatti possibile ordinare a posteriori i dati in modo da renderli crescenti, qualora questo fosse necessario. Tuttavia, operare nelle condizioni descritte comporta inevitabilmente una riduzione della “densità” dei punti acquisiti, cioè, per intenderci, la distanza in d.d.p. tra due acquisizioni successive di V_1 sarebbe maggiore di quanto ottenibile con una regolazione ottimale dei tempi.
- [9] Se ne avete voglia, potete comunque provare un best-fit per una curva che mostri la transizione tra regime di saturazione e regime attivo. Potete verificare facilmente che l'Eq. 3, opportunamente rimanipolata per V_{BE} co-

stante, conduce a $I_C = A - B \exp(-V_{CE}/(\eta V_T))$, con $A \simeq B \simeq \beta_F I_B$, dove β_F è il guadagno in corrente (per correnti continue) del transistor, qui supposto costante. Questa funzione descrive i dati in maniera non completa, ma tuttavia ragionevole, considerando le varie approssimazioni del modello di Ebers-Moll.

- [10] La regolazione della corrente di base tramite potenziometro rende difficile fare aggiustamenti fini. In una serie di dati non riportata in questa nota sono state eseguite misure per valori di I_B distanti tra loro di 1 μA (sempre nel range 10 – 15 μA) e i valori di I_C corrispondenti a $V_{CE} = 1.0$ V sono stati fittati a una retta passante per l'origine: la pendenza di questa retta corrisponde a β_f . Il valore ottenuto dal best-fit è in accordo con quello determinato nel modo, molto più immediato e grossolano, discusso nel testo.

Feedback e oscillatore a reazione (con transistor)

francesco.fuso@unipi.it

(Dated: version 6.b - Francesco Fuso, 27 febbraio 2022)

Questa nota ha il duplice scopo di esaminare alcuni aspetti generali legati al concetto di *feedback* in circuiti con transistor BJT e di vederne gli effetti in alcune realizzazioni pratiche in cui si instaura un regime di feedback negativo o positivo. In particolare la nota illustra realizzazione, principio di funzionamento e risultati ottenuti nell'esercitazione su condensatore di feedback e oscillatore a reazione basato su transistor a emettitore comune con rete di sfasamento.

I. INTRODUZIONE

Il concetto di *feedback* (o retro-azione) si applica nell'interpretazione di un'infinità di processi. Per esempio meccanismi di feedback si instaurano molto spesso in processi di interesse biologico e biomolecolare, regolandone l'evoluzione temporale.

In termini molto generali si realizza un feedback ogni qualvolta un sistema "sente" una data grandezza e aggiusta un qualche parametro applicando una certa *logica di controllo*. In un banalissimo esempio il sistema siete voi che vi fate una doccia: il *sensore* è la vostra pelle, che sente la temperatura dell'acqua, l'*attuatore* è la vostra manina che reagisce regolando la manopola del miscelatore, la logica di controllo vuole che la temperatura dell'acqua rimanga costantemente a un valore confortevole. Mentre fate la doccia può sicuramente verificarsi che sentiate l'acqua troppo fredda: dunque agirete sul miscelatore per aumentarne la temperatura, viceversa se l'acqua vi sembra troppo calda. In questo esempio sensore e attuatore sono due distinte funzionalità e la logica di controllo è quella, altamente sofisticata, del vostro cervello. Tuttavia esistono meccanismi di feedback in cui sensore e attuatore sono in qualche modo uniti nella funzionalità di un singolo elemento. Questo è il caso dei meccanismi che si possono realizzare con un transistor, un componente che, grazie alle propria natura attiva, si presta a integrare sensore e attuatore.

Abbiamo già discusso in un'altra nota la possibilità di realizzare feedback in un amplificatore a emettitore comune, ad esempio aggiungendo una resistenza nel ramo dell'emettitore, e abbiamo notato come questa possibilità recasse alcuni vantaggi pratici: l'aumento in stabilità delle condizioni di operazione, per quello che riguarda tensioni e correnti di polarizzazione, e l'accresciuta immunità da fenomeni di distorsione di non linearità, per quanto riguarda il comportamento con piccoli segnali variabili dal tempo. Abbiamo anche sottolineato come un effetto del feedback fosse quello di ridurre il guadagno dell'amplificatore: in termini di classificazione, questo suggeriva che il feedback fosse di tipo *negativo*.

II. CONDENSATORE IN PARALLELO A R_E (FILTRO PASSA-ALTO)

Ripetiamo brevemente i passaggi che, in una nota precedente, ci avevano condotto a esprimere il guadagno di un amplificatore a emettitore comune con resistenza di emettitore R_E del tipo rappresentato in Fig. 1(a). Nel modello introduciamo numerose approssimazioni, quali, ad esempio: (i) trascuriamo gli effettivi dettagli di funzionamento del transistor, in particolare la presenza di capacità parassite nelle giunzioni e l'eventuale passaggio di corrente attraverso giunzioni polarizzate inversamente; (ii) trascuriamo la possibilità che correnti variabili nel tempo possano fluire nei rami dedicati alla polarizzazione, in particolare quello per la polarizzazione della giunzione BE, o nel circuito di ingresso, in particolare nel generatore di forme d'onda; (iii) trascuriamo gli effetti dovuti alle impedenze finite del generatore di funzioni in ingresso (o generatore e partitore, come nell'esercitazione pratica), alla resistenza interna del generatore di d.d.p. continua usato per alimentare il circuito, all'impedenza di ingresso dell'oscilloscopio.

Supponendo che una corrente variabile nel tempo di ampiezza i_c entri nel collettore, ai capi di R_E si crea una piccola d.d.p. variabile nel tempo di ampiezza $R_E i_e$ (nello stesso verso di i_c rispetto alla maglia che comprende le giunzioni BC e BE, dunque con lo stesso segno), che dipende dall'ampiezza i_e della corrente variabile nel tempo che fluisce dall'emettitore. Nello scrivere l'equazione della maglia che comprende la giunzione BE occorre tenere conto di questa d.d.p.; in altri termini, viene modificata l'impedenza di ingresso del circuito e, di conseguenza, il guadagno. Le equazioni necessarie per il calcolo nell'ambito del modello che stiamo utilizzando sono riportate qui nel seguente:

$$\begin{aligned} i_e &= -\frac{i_c}{\alpha_f} = -(1 + \beta_f)i_b \\ v_{out} &= -R_C i_c = -R_C \beta_f i_b \\ v_{in} &= (R_B + r_b)i_b - R_E i_e = \\ &= [R_B + r_b + (1 + \beta_f)R_E]i_b \\ A_{V,RE} &= \frac{v_{out}}{v_{in}} = -\beta_f \frac{R_C}{R_B + r_b + (1 + \beta_f)R_E}, \end{aligned} \quad (1)$$

dove, al solito, v_{in} , v_{out} , i_b , i_c sono le ampiezze delle d.d.p. variabili nel tempo in ingresso e uscita e delle cor-

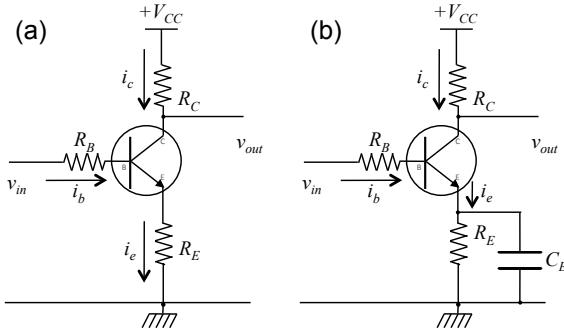


Figura 1. Amplificatore a emettitore comune con resistenza di emettitore R_E (a) e con l'ulteriore aggiunta di un condensatore in parallelo C_E (b). Le parti di circuito dedicate alla polarizzazione delle giunzioni non sono indicate: $+V_{CC}$ rappresenta il polo positivo dell'alimentazione.

renti variabili nel tempo che scorrono nel collettore e nell'emettitore, e $\beta_f = \alpha_f / (\alpha_f - 1)$ è il guadagno in corrente per piccoli segnali variabili nel tempo (analogamente α_f è il coefficiente dell'effetto transistor in queste condizioni). Notate che, per rigore e compatibilità con altre descrizioni, nell'espressione di v_{in} si è posto un segno negativo davanti al termine $R_E i_e$ allo scopo di soddisfare quanto affermato prima, cioè che la d.d.p. variabile nel tempo che si forma ai capi di R_E è positiva quando la corrente scorre “dall'alto verso il basso” nel ramo che comprende collettore ed emettitore [guardate Fig. 1(a)]. Il guadagno con feedback $A_{V,RE}$ risulta diminuito rispetto alla configurazione imperturbata, cioè senza R_E , circostanza che ci permette di identificare come negativo il meccanismo di feedback realizzato.

Aggiungiamo un condensatore C_E in parallelo a R_E come nello schema di Fig. 1(b). Restringendoci a esaminare *segnali sinusoidali* (in ingresso, e quindi per tutte le d.d.p. e correnti considerate), la presenza del condensatore implica che la d.d.p. variabile nel tempo ai capi del parallelo sia esprimibile come il fasore $Z_{eq} i_{\omega,e}$, con Z_{eq} impedenza del parallelo R_E con C_E :

$$Z_{eq} = \frac{1}{1/R_E + j\omega C_E} = \frac{R_E}{1 + j\omega R_E C_E}. \quad (2)$$

Operando con il metodo simbolico, il guadagno in tensione espresso in Eq. 1 diventa una *funzione di trasferimento*:

$$\begin{aligned} T_{RE,CE} &= -\frac{v_{\omega,out}}{v_{\omega,in}} = \\ &= -\beta_f \frac{R_C}{R_B + r_b + (1 + \beta_f)R_E \frac{1}{1 + j\omega R_E C_E}}; \end{aligned} \quad (3)$$

essa dà luogo al guadagno $A_{V,RE,CE} = |T_{RE,CE}|$ (non si riporta l'espressione analitica) che è evidentemente funzione della frequenza angolare ω . Infatti l'impedenza del parallelo $R_E // C_E$ tende a zero per frequenze alte dove, di conseguenza, gli effetti negativi del feedback tendono ad annullarsi.

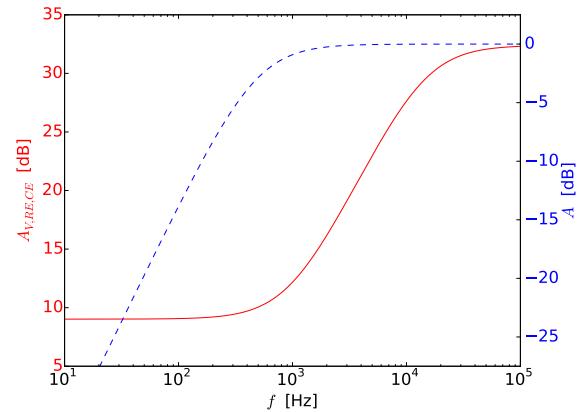


Figura 2. Diagramma di Bode del circuito di Fig. 1(b) (linea rossa continua, asse verticale di sinistra) calcolato, nell'ambito del modello discusso nel testo, usando Eq. 3. Per riferimento lo stesso grafico riporta anche il guadagno A_V di un filtro passa-alto passivo a un polo (linea tratteggiata blu, asse verticale di destra) realizzato con gli stessi valori di R_E e C_E impiegati nel circuito di Fig. 1(b) ($R_E = 0.33$ kohm, $C_E = 1.0 \mu\text{F}$).

La Fig. 2 mostra il diagramma di Bode (linea rossa continua) corrispondente alla funzione di trasferimento di Eq. 3, calcolato supponendo un guadagno imperturbato (in assenza sia di C_E che di R_E) $A_V \approx 32$ dB, con $R_E = 0.33$ kohm, $C_E = 1.0 \mu\text{F}$. Per referenza, lo stesso grafico riporta anche il guadagno di un filtro passa-alto passivo a un polo (linea blu tratteggiata) con frequenza angolare di taglio $\omega_T = 1/(R_E C_E)$: si osserva come la risposta sia diversa, in particolare come il guadagno tenda a un valore finito per basse frequenze e come le frequenze caratteristiche (per esempio la *corner frequency*) siano diverse nei due casi, a causa della specifica funzione di trasferimento trovata, in particolare per la presenza del coefficiente β_f che si trova a moltiplicare la parte dipendente dalla frequenza.

Dal punto di vista sperimentale il modello applicato può rivelarsi inadeguato, in particolare per la mancata considerazione delle capacità di giunzione presenti nel transistor, le quali, fra gli altri effetti, sono responsabili della diminuzione del guadagno a frequenze alte attraverso un fenomeno operativamente simile a quello illustrato in Sez. III.

III. CONDENSATORE DI FEEDBACK (FILTRONE PASSA-BASSO)

In termini molto generali, avere un meccanismo di feedback richiede di “re-inviare” parte del segnale di uscita all’ingresso attraverso un opportuno “condizionamento”. Nell’esempio della doccia, l’“uscita” rappresenta la temperatura dell’acqua, l’“ingresso” è la regolazione del miscelatore, il “condizionamento” è la logica con la quale

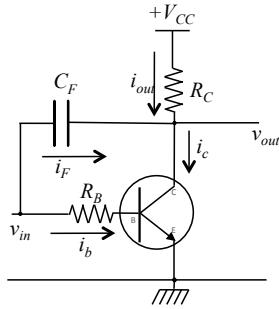


Figura 3. Amplificatore a emettitore comune con condensatore di feedback C_F . Le parti di circuito dedicate alla polarizzazione delle giunzioni non sono indicate: $+V_{CC}$ rappresenta il polo positivo dell'alimentazione.

viene comandata la manopola del miscelatore in funzione del confort corporale ricercato.

Anche nell'esempio della resistenza di emettitore, senza o con condensatore in parallelo, una parte dell'uscita, il segnale i_c , ovvero i_e , viene trasferita in ingresso attraverso la conseguente modifica di i_b (a parità di v_{in}). In quell'esempio il metodo con cui il trasferimento viene realizzato è piuttosto complicato; un metodo più diretto consiste nel prelevare fisicamente parte del segnale di uscita e riportarlo all'ingresso attraverso un opportuno ramo. È evidente che questo ramo deve operare solo sulle componenti variabili nel tempo dei segnali, perché altrimenti le condizioni di polarizzazione del transistor, che vogliamo sempre in regime attivo, verrebbero modificate in maniera potenzialmente incongrua.

Il modo più semplice per realizzare questo tipo di feedback è costituito da un condensatore C_F che collega il collettore alla base (alla resistenza R_B per le esigenze pratiche dei componenti che abbiamo a disposizione) secondo lo schema di Fig. 3: infatti il condensatore, comportandosi da circuito aperto per le correnti continue, non modifica le condizioni di polarizzazione delle giunzioni. L'aggiunta del condensatore crea un ramo che è attraversato da una corrente variabile nel tempo di ampiezza i_F e determina un nodo al collettore; detta i_{out} l'ampiezza della corrente variabile nel tempo che attraversa R_C , vale la relazione $i_c = i_{out} + i_F$. Osservate che i segni delle correnti variano nel tempo, in particolare per segnali periodici alternati, e che il segno di i_F dipende dalla differenza ($v_{out} - v_{in}$): nella figura si suppone che essa sia negativa.

Applicando le approssimazioni citate in Sez. II e usando il metodo simbolico per segnali sinusoidali, il modello fornisce le seguenti equazioni:

$$\begin{aligned} i_{\omega,F} &= \frac{v_{\omega,in} - v_{\omega,out}}{Z_F} \\ i_{\omega,c} &= i_{\omega,out} + i_{\omega,F} \\ v_{\omega,out} &= -R_C i_{\omega,out} \\ v_{in} &= (R_B + r_b) i_{\omega,b} \\ i_{\omega,c} &= \beta_f i_{\omega,b} \end{aligned} \quad (4)$$

Combinando le equazioni si ottiene

$$\begin{aligned} i_{\omega,out} &= -\frac{v_{\omega,out}}{R_C} = i_{\omega,c} - i_{\omega,F} = \\ &= \beta_f \frac{v_{in}}{R_B + r_b} - \frac{v_{in}}{Z_F} + \frac{v_{out}}{Z_F}, \end{aligned} \quad (5)$$

da cui

$$v_{out} \left(\frac{1}{Z_F} + \frac{1}{R_C} \right) = v_{in} \left(\frac{1}{Z_F} - \frac{\beta_f}{R_B + r_b} \right). \quad (6)$$

Quindi è

$$T_{ZF} = \frac{v_{\omega,out}}{v_{\omega,in}} = \frac{R_C}{R_C + Z_F} \frac{R_B + r_b - \beta_f Z_F}{R_B + r_b}. \quad (7)$$

Restringendoci al caso di interesse, in cui $Z_F = 1/(j\omega C_F)$, ulteriori semplici manipolazioni di Eq. 7 portano a

$$T_{CF} = -\beta_f \frac{R_C}{R_B + r_b} \frac{1 - \frac{j\omega(R_B + r_b)C_F}{\beta_f}}{1 + j\omega R_C C_F}. \quad (8)$$

Questa funzione di trasferimento differisce da quella di un ordinario filtro passa-basso (passivo, a un polo) per la presenza del termine immaginario al numeratore; questo termine compare tuttavia diviso per β_f , per cui il suo contributo è piccolo, almeno a frequenze angolari non troppo elevate. La Fig. 4 rappresenta il diagramma di Bode del guadagno $A_{V,CF} = |T_{CF}|$ determinato da Eq. 8 in funzione della frequenza (linea rossa continua) assieme, come in Fig. 1, alla curva di riferimento per un ordinario filtro passa-basso passivo a un polo con frequenza angolare di taglio $\omega_T = 1/(R_C C_F)$.

L'effetto del feedback, in particolare il suo carattere negativo, può essere interpretato qualitativamente: nel caso di un amplificatore *invertente* come quello a transistor a emettitore comune, riportare parte del segnale in uscita all'ingresso (in realtà con uno sfasamento di $\pi/2$ rad tra correnti e tensioni) determina una *diminuzione* del guadagno, tanto più marcata quanto minore, in modulo, è l'impedenza del ramo di feedback. Dunque all'aumentare della frequenza il feedback diviene sempre più negativo e il guadagno del circuito diminuisce sempre di più rispetto al circuito imperturbato.

Un'altra conseguenza della presenza di C_F nel circuito è la modifica dell'impedenza di ingresso Z_{in} . In condizioni imperturbate (senza C_F) il nostro modello porta a $Z_{in} = (R_B + r_b)$. Il ramo di feedback, però, comporta la presenza di un nodo prima di R_B in cui la debole corrente variabile nel tempo si dirama. Il ramo che comprende Z_F prosegue poi con un ulteriore nodo, da cui la corrente dirama in una parte che va al collettore e una parte che passa attraverso R_C . Supponendo di operare a frequenze tali da rendere sufficientemente piccola $|Z_F|$ rispetto a $(R_B + r_b)$, la presenza di questi nodi viene al pettine. In queste condizioni, come abbiamo stabilito, il guadagno tende a zero, per cui le correnti i_b e i_c diventano piccole. Pertanto la maggior parte della corrente associata al segnale di ingresso v_{in} si trova a fluire per la serie $Z_F + R_C$

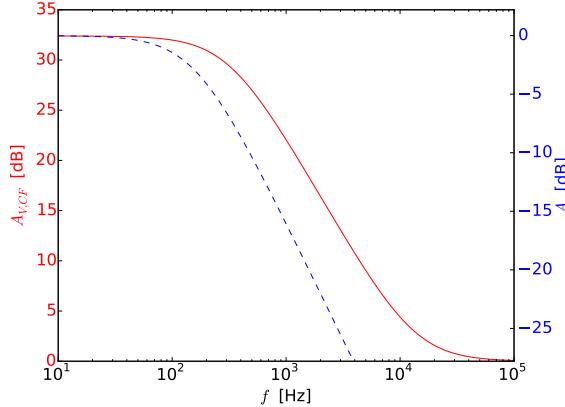


Figura 4. Guadagno del circuito di Fig. 3 (linea rossa continua, asse verticale di sinistra) calcolato, nell’ambito del modello discusso nel testo, usando Eq. 8. Per riferimento lo stesso grafico riporta anche il guadagno A_V di un filtro passa-basso passivo a un polo (linea tratteggiata blu, asse verticale di destra) realizzato con gli stessi valori di R_C e C_F impiegati nel circuito di Fig. 3 ($R_C = 1.0$ kohm, $C_F = 1.0 \mu\text{F}$). Nel calcolo si è supposto un guadagno imperturbato (senza C_F) $A_V \simeq 32$ dB.

(la maglia si richiude attraverso il generatore V_{CC} e le due impedanze sono in serie perché abbiamo supposto $i_c \rightarrow 0$), che tende quindi a rappresentare l’impedenza di ingresso. Nell’esercitazione pratica, dove il segnale v_{in} è fornito dal generatore di funzioni accoppiato al partitore di tensione con condensatore, che risulta in un generatore reale con impedenza di Thévenin di modulo relativamente alto ($|Z_{Th}| \simeq 2.7$ kohm a 1 kHz), l’aggiunta al circuito del condensatore C_F può portare a una vistosa attenuazione del segnale v_{in} osservato all’oscilloscopio, fino al punto di impedire misure. Per aumentare l’ampiezza del segnale può essere sufficiente rimuovere l’effetto di partizione, scambiando tra loro le resistenze di cui è dotato il partitore.

IV. FEEDBACK LOOP

In termini molto generali un amplificatore con feedback può essere rappresentato come in Fig. 5: il ramo di feedback, che ha un effetto generalmente dipendente dalla frequenza rappresentato da una generica funzione di trasferimento ϵ , collega ingresso e uscita dell’amplificatore il cui guadagno imperturbato è $A < 0$ (amplificatore invertente, supponiamo A reale, cioè trascuriamo eventuali effetti di sfasamento a carico dell’amplificatore imperturbato): in questo modo si realizza un *feedback loop*.

Un’applicazione di questo approccio, in cui si trascognano tutti i dettagli relativi al funzionamento effettivo dell’amplificatore e si applicano tutte le approssimazioni elencate in Sez. II, è la seguente. Supponendo segnali

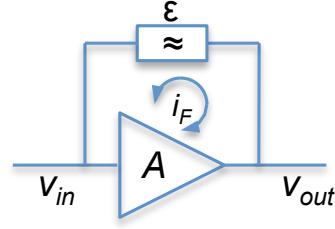


Figura 5. Illustrazione del feedback loop con amplificatore.

sinusoidali, e quindi usando il metodo simbolico, si ha

$$v_{\omega,out} = A(v_{\omega,in} + \epsilon v_{\omega,out}), \quad (9)$$

da cui il guadagno con feedback, o in *closed loop*:

$$A_F = \left| \frac{v_{\omega,out}}{v_{\omega,in}} \right| = \frac{|A|}{|1 - A\epsilon|}. \quad (10)$$

Assumendo per ulteriore semplificazione $\epsilon = B$ reale e positivo, cioè assenza di sfasamenti, il guadagno con feedback diventa

$$A_F = \frac{|A|}{1 - AB}; \quad (11)$$

poiché è $A < 0$, l’Eq. 11 indica che il guadagno con feedback (in closed loop) si riduce in valore assoluto rispetto a quello imperturbato (in *open loop*), cioè il feedback è negativo.

Una conseguenza di questo approccio, che è molto utile nell’impiego degli amplificatori operazionali che vedrete nel prossimo anno, è un effetto che viene spesso chiamato *effetto Miller*. Per verificarlo dobbiamo supporre che il nostro amplificatore assorba una corrente molto bassa, virtualmente nulla, ovvero $i_b \ll i_F$. In queste condizioni la corrente di feedback si ritrova inalterata come corrente i_{in} nella maglia di ingresso (a cui sarà collegato un qualche circuito, per esempio il generatore di funzioni). Per l’impedenza di ingresso, si ha allora $Z_{in} = v_{in}/i_{in} = v_{in}/i_F$. Immaginando che l’anello di feedback sia costituito da un semplice condensatore C_F e che i segnali siano sinusoidali, quindi $Z_F = 1/(j\omega C_F)$, è $i_{\omega,F} = j\omega C_F(v_{\omega,out} - v_{\omega,in})$, da cui $Z_{in} = Z_F/(A - 1)$: se $|A| \gg 1$, l’impedenza di ingresso diventa, in modulo, quella di un condensatore di capacità A -volte più grande.

Questo effetto, simile a quanto si verificava per la resistenza di emettitore R_E in Sez. II, che nell’impedenza di ingresso era vista come se fosse moltiplicata per un fattore simile a β_f , gioca un ruolo nel limitare la banda passante degli amplificatori con transistor BJT. Infatti all’interno delle loro giunzioni, in particolare della BC, si determinano delle capacità spurie, che, venendo “amplificate” dall’effetto Miller, danno luogo a un comportamento spurio di tipo filtro passa-basso in analogia con quanto abbiamo discusso in Sez. III.

V. FEEDBACK POSITIVO

Supponiamo ora che l'anello di feedback abbia una funzione di trasferimento ϵ tale da creare uno sfasamento di $\pm\pi$ rad tra v_{out} e v_{in} . In altre parole, il segnale re-invito in ingresso ha segno opposto (ed è attenuato in valore assoluto, supponendo il ramo di feedback sia realizzato con componenti passivi) rispetto a quello che si trova in ingresso. In queste condizioni, si ha in sostanza $\epsilon = B$ reale, con $-1 < B < 0$ e il denominatore di Eq. 11 può annullarsi se $AB = 1$. La conseguenza è che il guadagno in closed-loop tende a divergere, cioè che l'ampiezza del segnale in uscita tende a divergere. Il feedback così realizzato può essere definito *positivo*.

Tale divergenza è chiaramente non fisica: l'ampiezza divergente implica una potenza fornita al circuito che può aumentare all'infinito e, inoltre, l'ampiezza del segnale in uscita è limitata dalla d.d.p. finita fornita dal generatore che alimenta il circuito. Come vedremo nel seguito, la fisica può costringere il sistema ad *auto-oscillare* invece che assumere un guadagno infinito, cioè il segnale in uscita diventa una funzione (quasi-)periodica che si auto-sostenta prelevando potenza dal generatore.

Un esempio degli effetti eclatanti del feedback positivo familiare a molti è il cosiddetto *effetto Larsen*. Immaginate di avere in mano un microfono collegato a un impianto di amplificazione audio e a degli altoparlanti: è possibile che, orientando il microfono verso gli altoparlanti, si innesci una oscillazione spontanea e si senta un fischio, che rappresenta appunto l'auto-oscillazione del sistema. Per disinnescarla è in genere sufficiente indirizzare il microfono altrove oppure ridurre il guadagno, cioè agire sulla manopola del volume dell'amplificatore. Un esempio di importanza fondamentale, a cui accenneremo al termine del corso, è il laser: la lettera "a" dell'acronimo sta proprio per amplificatore e la presenza di opportuni meccanismi di feedback positivo, che agiscono sul numero di fotoni, permette di ottenere radiazione elettromagnetica che oscilla a una determinata frequenza. Un esempio storico è il famoso Tacoma Narrows Bridge, che crollò proprio per l'innescarsi di auto-oscillazioni e non per il fenomeno della risonanza come qualche volta riportato nei libri di fisica delle scuole: la risonanza è infatti un fenomeno che ha origini, leggi e rappresentazioni matematiche diverse.

Prima di aggiungere altri dettagli concettuali e, soprattutto, discutere gli aspetti pratici legati alla realizzazione sperimentale di un auto-oscillatore a transistor BJT, torniamo all'esempio del feedback realizzato sotto la doccia: per esperienza personale sappiamo tutti che è possibile che si verifichino delle oscillazioni nella temperatura dell'acqua. Questo si verifica per la presenza di due ingredienti principali: l'esistenza di un *ritardo* tra regolazione del miscelatore ed effettiva variazione di temperatura dell'acqua (c'è un tratto di tubo in mezzo) e/o la presenza di *non linearità* (ruotare la manopola del miscelatore può dare effetti diversi a riscaldare o a raffreddare e certamente la temperatura non può scen-

dere sotto o salire sopra determinati valori stabiliti dalla caldaia). Entrambi questi ingredienti sono generalmente presenti in un auto-oscillatore, detto anche *oscillatore a reazione*, con transistor BJT. Infatti la presenza di un ritardo è intrinseca nello sfasamento introdotto nell'anello di feedback, le non linearità sono tipiche nell'impiego di un amplificatore a transistor, in particolare nella configurazione a emettitore comune, così come la presenza di vincoli nell'ampiezza di uscita.

A. Ritardo e auto-oscillazione

La matematica che serve per descrivere l'auto-oscillazione è al di fuori della portata del secondo anno di corso di Fisica. Inoltre il modello matematico specifico dell'oscillatore a reazione costruito nell'esercitazione pratica è estremamente complicato per la presenza di diverse maglie e molto poco maneggevole.

Tuttavia, facendo sempre riferimento a un modello semplificato e approssimativo come quello di Fig. 5, possiamo individuare una possibile origine matematica per l'auto-oscillazione eseguendo un'analisi nel dominio dei tempi: in questa analisi i segnali di interesse saranno considerati come funzione del tempo (e nei simboli useremo, per esempio, $v_{out}(t)$ per ricordarcelo). Una lettura utile per illustrare l'approccio in altri settori della fisica è A. Jenkins, "Self-oscillation", disponibile in arxiv.org.

Per prima cosa osserviamo che lo sfasamento di $\pm\pi$ rad realizzato dalla funzione di trasferimento ϵ nell'anello di feedback si traduce, nel dominio dei tempi, in un certo ritardo temporale τ , che qui non ci preoccupiamo di determinare, fisicamente corrispondente al tempo necessario affinché una variazione del segnale possa essere trasferita dall'uscita all'ingresso. Allora all'istante t l'anello di feedback re-invia all'ingresso una frazione dell'uscita all'istante $t - \tau$.

In analogia con Eq. 9, possiamo allora scrivere

$$v_{out}(t) = A(v_{in}(t) + Bv_{out}(t - \tau)), \quad (12)$$

dove $-1 < B < 0$ tiene conto di sfasamento e attenuazione. Il termine $v_{out}(t - \tau)$ può essere sviluppato al secondo ordine attorno a $\tau = 0$ [cioè per $(t - \tau) \rightarrow t$], ottenendo

$$\begin{aligned} v_{out}(t) = & Av_{in}(t) + ABv_{out}(t) - AB\tau \frac{dv_{out}(t)}{dt} + \\ & + AB \frac{\tau^2}{2} \frac{d^2v_{out}(t)}{dt^2}. \end{aligned} \quad (13)$$

Notate che non ci vogliamo qui preoccupare della validità dello sviluppo, che in generale non sapremmo giudicare non potendo stimare alcuno dei termini che vi compaiono, ma solo sviluppare al secondo ordine l'Eq. 12 allo scopo di verificare se essa ammette soluzioni oscillanti di tipo armonico.

L'Eq. 13 può essere riscritta come

$$\frac{d^2v_{out}(t)}{dt^2} - \frac{2}{\tau} \frac{dv_{out}(t)}{dt} + \left(\frac{2}{\tau^2} \left(1 - \frac{1}{AB} \right) \right) v_{out}(t) = -\frac{2}{B\tau^2} v_{in}(t), \quad (14)$$

che ha una qualche analogia formale con l'equazione di un oscillatore armonico forzato, purché $AB > 1$ (che supponiamo verificata), ma con la rilevantissima differenza del *segno* del termine che moltiplica la derivata prima. In meccanica questo termine tiene conto dello smorzamento viscoso a cui corrisponde una forza di attrito, per cui il termine è positivo. Qui è come se lo smorzamento agisse in “senso opposto”, facendo aumentare, invece che diminuire, l'ampiezza delle oscillazioni. Se ricordate il ruolo della forzante in un oscillatore meccanico, necessaria per ripristinare l'energia perduta tramite attrito, potrete concludere che, in presenza di un “anti-smorzamento”, della forzante non c'è più bisogno. In altre parole, il secondo membro può essere azzerato, la v_{in} non serve più e il sistema *auto*-oscilla. Chiaramente le auto-oscillazioni devono ricevere energia da qualche sorgente, e poiché l'energia è sempre necessariamente finita, l'andamento esponenziale crescente della soluzione di Eq. 14 non è fisico, come abbiamo già osservato discutendo l'amplificazione in closed loop. In realtà, infatti, il nostro modello ha trascurato la circostanza che il ramo di feedback è passivo e quindi, nell'attenuare il segnale che lo attraversa, dissipava energia, o potenza. Un'equazione più realistica, ma anche più generale, potrebbe essere:

$$\frac{d^2v_{out}(t)}{dt^2} - (\gamma_+ - \gamma_-) \frac{dv_{out}(t)}{dt} + \omega_{auto}^2 v_{out}(t) = 0, \quad (15)$$

dove abbiamo introdotto dei coefficienti di anti-smorzamento e smorzamento (rispettivamente γ_+ e γ_- , entrambi positivi) e indicato con ω_{auto} la frequenza angolare di auto-oscillazione. Naturalmente il valore effettivo che questi parametri assumono dipende dallo specifico sistema considerato: ad esempio, nell'oscillatore a reazione con transistor di cui ci occuperemo nel seguito ω_{auto} è prossima alla frequenza angolare che determina lo sfasamento di $\pm\pi$ rad richiesto per il feedback positivo. L'espressione di Eq. 15 è utile perché suggerisce immediatamente che le auto-oscillazioni possono verificarsi solo se $\gamma_+ > \gamma_-$, cioè, usando una terminologia comune in questo ambito, se *il guadagno del feedback loop prevale sulle perdite*.

Restano due aspetti da chiarire. Affinché l'oscillatore possa avviarsi è necessario fornire delle condizioni iniziali, cioè un *innesco* per il meccanismo di feedback positivo. Come illustreremo nel seguito, normalmente per l'innesco basta una fluttuazione del segnale in ingresso, per esempio dovuta a rumore. Infine la soluzione di Eq. 15 è armonica perché abbiamo troncato lo sviluppo di $v_{out}(t-\tau)$ al secondo ordine. La presenza rilevante di ordini superiori può condurre ad auto-oscillazioni anarmoniche, che sono anche tipiche in sistemi di questo tipo.

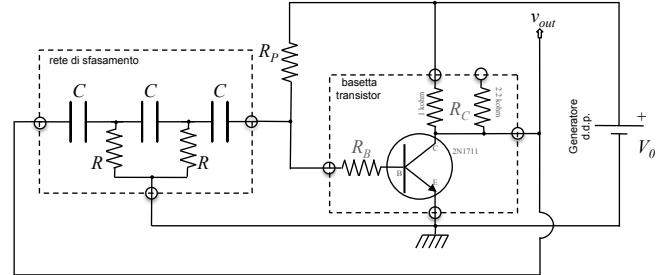


Figura 6. Circuito di oscillatore a reazione basato su transistor BJT (2N1711) montato in configurazione a emettitore comune e rete di sfasamento RC costituita da due resistenze e tre condensatori preassemblati in un telaietto con tre boccole.

VI. OSCILLATORE A REAZIONE E RETE DI SFASAMENTO

L'esercitazione pratica prevede di realizzare un (auto-)oscillatore a reazione in cui il ramo di feedback sfase le d.d.p. in ingresso rispetto all'uscita di π rad. Come rappresentato in Fig. 6 l'anello di feedback è costituito da una *rete di sfasamento RC* . Infatti, come ben sappiamo, circuiti RC (derivatori e integratori) realizzano uno sfasamento tra ingresso e uscita che, almeno nel caso di segnali sinusoidali, può essere facilmente modellato usando il metodo simbolico.

Sappiamo anche che esistono molte altre possibilità per ottenere uno sfasamento, per esempio usando (anche) elementi induttivi. Sicuramente la scelta di impiegare una rete di sfasamento basata solo su resistenze e capacità può non essere ottimale sotto molti punti di vista, ma essa ha l'indubbio vantaggio della semplicità concettuale e realizzativa e permette di ottenere con pochi sforzi un oscillatore a reazione funzionante.

La rete di sfasamento è costituita da due resistenze e tre condensatori ceramici preassemblati in un telaietto dotato di tre boccole. I valori nominali dei componenti sono $R = 3.0$ kohm e $C = 0.1 \mu F$, con tolleranze dell'ordine del $\pm 5\%$ e $\pm 10\%$ per resistenze e capacità (potrebbero essere montati componenti con tolleranze diverse in telaietti diversi).

Come si vede, i condensatori sono “in serie” rispetto al “segnale”: in queste condizioni essi possono comportarsi da derivatori, ovvero dare luogo a filtri passa-alto. Per un *singolo* derivatore RC , funzione di trasferimento $T(f)$, guadagno $A(f)$, sfasamento $\Delta\phi$ e frequenza di taglio f_T

sono dati da

$$T(f) = \frac{1}{1 - j f_T/f} \quad (16)$$

$$A(f) = \frac{1}{\sqrt{1 + (f_T/f)^2}} \quad (17)$$

$$\tan(\Delta\phi) = \frac{f_T}{f} \quad (18)$$

$$f_T = \frac{1}{2\pi R C}, \quad (19)$$

dove si è ovviamente supposto di operare con segnali sinusoidali. Il valore nominale della frequenza di taglio è $f_T \simeq 0.53$ kHz dato da Eq. 19.

È evidente da Eq. 18 che lo sfasamento prodotto da un singolo derivatore vale al massimo $\pi/2$ rad (per $f \rightarrow 0$) e dunque con un singolo derivatore (o con un circuito che faccia uso di una sola resistenza e un solo condensatore) non si riesce a conseguire l'obiettivo di invertire il segnale. Nell'approssimazione pesante, ma utile, in cui si suppone che nella cascata lo stadio a valle perturbi in maniera trascurabile quello a monte, lo sfasamento complessivo è dato dalla somma dei singoli sfasamenti introdotti da ogni stadio. Dunque già con soli due derivatori in cascata potrebbe essere raggiunto lo sfasamento complessivo di π rad. Tuttavia ogni derivatore dovrebbe produrre uno sfasamento di $\pi/2$ rad, e quindi lavorare in condizioni di attenuazione molto elevata [$A(f) \rightarrow 0$ per $f \rightarrow 0$] che potrebbero non consentire la condizione, citata prima, che il guadagno in closed-loop prevalga sulle perdite. Per questo motivo è sicuramente preferibile prevedere l'uso di (almeno) tre derivatori in cascata. Usando ancora l'approssimazione di derivatori indipendenti e supponendoli identici tra loro, cioè supponendo la presenza di una terza resistenza tra l'ultimo condensatore e linea di massa, o terra, ognuno dovrebbe sfasare di $\pi/3$, per cui la frequenza di operazione sarebbe $f = f_T/\sqrt{3} \simeq 0.31$ kHz.

Tuttavia questa terza resistenza vedrebbe la resistenza di ingresso R_{in} dell'amplificatore in parallelo e avrebbe la poco piacevole conseguenza di alterare le condizioni di polarizzazione della giunzione BE. Pertanto essa non è di fatto presente nella rete di sfasamento, riportata per chiarezza in Fig. 7, che può essere vista come la composizione di tre derivatori in cascata di cui l'ultimo vede come resistenza un'impedenza di carico Z_L che può approssimativamente essere identificata nella resistenza di ingresso dell'amplificatore a emettitore comune, cioè $R_{in} = R_B + r_b$, dove nell'ultimo passaggio abbiamo supposto la resistenza di polarizzazione, R_P in Fig. 6, sufficientemente grande da impedire un significativo passaggio di corrente nel ramo in cui è inserita. Inoltre abbiamo evidentemente trascurato (come al solito!) le capacità spurious presenti nel transistor. Infine il modello approssimativo prevede che il generatore di segnale (sinusoidale) collegato all'ingresso della rete di sfasamento sia ideale, cioè dotato di resistenza trascurabile. Nella configurazione di Fig. 6 tale resistenza può essere identificata con la resistenza di uscita dell'amplificatore a emettitore co-

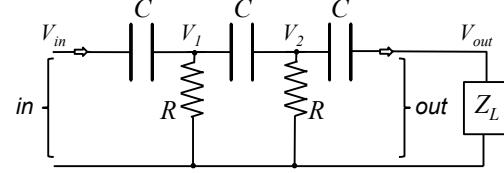


Figura 7. Schema della rete di sfasamento con indicazione di ingresso e uscita e dell'“impedenza di carico” Z_L , rappresentata da un rettangolo; sono anche riportati i punti, o nodi, in cui si misurano diverse d.d.p., tutte riferite alla linea di massa, come citato nel testo.

mune, per il quale sappiamo che, approssimativamente, $R_{out} = R_C$. Il valore di R_C non è piccolo in termini assoluti, per cui torneremo a posteriori a valutare questa affermazione.

Supponendo che la configurazione a cascata non modifichi le caratteristiche dei singoli stadi della rete di sfasamento (fate riferimento a Fig. 7 per individuare le d.d.p. rilevanti) si ha

$$\begin{aligned} V_{\omega,1} &= V_{\omega,in} T_1 = V_{\omega,in} \frac{1}{1 - j f_T/f} \\ V_{\omega,2} &= V_{\omega,1} T_2 = V_{\omega,1} \frac{1}{1 - j f_T/f} \\ V_{\omega,out} &= V_{\omega,2} T_3 = V_{\omega,2} \frac{1}{1 - j f_{TL}/f}, \end{aligned} \quad (20)$$

con $f_{TL} = 1/(2\pi R_{in}C)$ frequenza di taglio dell'“ultimo” derivatore.

Le d.d.p. indicate in Eq. possono essere effettivamente misurate nell'esperimento, dove la funzionalità della rete di sfasamento può essere analizzata usando il generatore di forme d'onda, impostato per una forma sinusoidale, e l'oscilloscopio. Osservate che, in questo caso, nell'analisi di V_{out} occorre tenere conto della resistenza di ingresso dell'oscilloscopio, $R_{osc} = 1$ Mohm (nominale), che di fatto costituisce l'ultima resistenza della cascata di derivatori; questo comporta che lo sfasamento di π rad si abbia a frequenze attorno a 20 Hz.

A. Equazione completa della rete di sfasamento

Il modello della rete di sfasamento da cui siamo partiti, basato sull'indipendenza degli stadi, può essere ritenuto corretto solo in una forma molto approssimativa, che, appunto, serve per considerazioni grossolane (all'ordine di grandezza). Infatti i tre derivatori hanno impedenze uguali (o simili, per l'ultimo) fra loro e quindi l'affermazione che il derivatore successivo non “perturbi” il precedente è poco ragionevole: occorrerebbe infatti che che lo stadio successivo avesse un'impedenza (in modulo) maggiore del precedente, come ben sappiamo.

Il problema di modellare adeguatamente la rete di sfasamento può essere risolto facilmente scrivendo le equa-

zioni nel metodo simbolico per le tre *maglie* che costituiscono l'intera rete di sfasamento e quindi imponendo che la funzione di trasferimento complessiva sia reale. Le equazioni delle tre maglie (si scelgono quelle comprendenti ognuna un condensatore, numerandole dall'inizio alla fine) potrebbero essere

$$V_{\omega,in} = (R + \frac{1}{j\omega C})I_{\omega,1} - RI_{\omega,2} \quad (21)$$

$$0 = (2R + \frac{1}{j\omega C})I_{\omega,2} - RI_{\omega,2} - RI_{\omega,3} \quad (22)$$

$$0 = (R + R' + \frac{1}{j\omega C})I_{\omega,3} - RI_{\omega,2}, \quad (23)$$

dove R' indica la resistenza “vista” dall’ultimo derivatore. Supponendo le tre resistenze di valore R uguale fra loro (cioè $R' = R$), la frequenza angolare ω_π in cui la funzione di trasferimento è reale risulta

$$\omega_\pi = \frac{1}{\sqrt{6RC}}, \quad (24)$$

che differisce per un fattore $1/\sqrt{2}$ rispetto a quella trovata supponendo indipendenti i tre derivatori. Nel caso in cui $R' \neq R$ si avrebbe invece

$$\omega_\pi = \frac{1}{\sqrt{3(R + R')RC^2}}, \quad (25)$$

Naturalmente le dimostrazioni si lasciano per esercizio!

Come già affermato, nelle condizioni dell'esperimento si può supporre $R' = R_{in}$, resistenza di ingresso dell'amplificatore a emettitore comune. Supponendo valida l'approssimazione per cui $r_b \simeq \eta V_T / I_B$, R_{in} dipende dalla corrente di base e quindi dal valore di R_P . Occorre tuttavia sottolineare come la regolazione del potenziometro possa essere tale da rendere tutt’altro che trascurabile la corrente variabile nel tempo che dirama attraverso R_P e da qui a terra attraverso la resistenza interna di Arduino. Inoltre, essendo la resistenza interna di Arduino non trascurabile (la corrente massima che una sua porta digitale può fornire è, secondo datasheet, 20 mA), può verificarsi che per valori di R_P sufficientemente bassi, corrispondenti a elevata richiesta di corrente (sia I_B che, soprattutto, $I_C = \beta_F I_B$), la d.d.p. erogata dalla sua porta digitale diminuisca, provocando una diminuzione della corrente di base I_B e un comportamento non monotono dell'intensità di corrente di polarizzazione con la posizione dell'alberino del potenziometro.

Infine, possiamo verificare a posteriori che l'effetto della resistenza finita, R_C , in uscita all'amplificatore sia trascurabile. In effetti, detta Z_{sf} l'impedenza complessiva della rete di sfasamento, si ha sempre $|Z_{sf}| > R_C$; in particolare, nelle condizioni operative di cui alla prossima sezione, si ha $|Z_{sf}|/R_C \gtrsim 10$, per cui, considerando l'accuratezza tipica delle nostre misure, l'approssimazione può essere ritenuta valida.

Come già affermato, l'auto-oscillazione ha bisogno di un innescio, o se preferite un “seme” di tensione, che, per

qualche motivo, deve presentarsi all’ingresso dell’amplificatore. Fluttuazioni che assomigliano a impulsi possono facilmente formarsi all’atto dell'accensione di un circuito, cioè nel transitorio che si verifica subito dopo che esso comincia ad essere alimentato. Infatti in queste condizioni da qualche parte del circuito (non solo l'amplificatore e la rete di sfasamento, ma anche l'alimentatore che si impiega) si possono formare degli impulsi di corrente, che facilmente possono accoppiarsi all’ingresso dell'amplificatore arrivando alla base del transistor sotto forma di impulsi di tensione.

Per esaminare in modo un po’ meno qualitativo come un impulso possa innescare un oscillazione periodica, o quasi-periodica, possiamo fare uso del concetto di trasformata di Fourier e analizzare il sistema nel dominio delle frequenze: a un impulso nel tempo può essere associato un segnale oscillante fatto dalla sovrapposizione di moltissime frequenze diverse (idealmente, un impulso rappresentato da una funzione delta ha uno spettro che comprende tutte le frequenze; realisticamente, lo spettro contiene un intervallo finito di frequenze che dipende da durata e forma dell’impulso). Tra tutte queste componenti a diverse frequenze potrà essercene una (potranno essercene alcune) per le quali la rete di sfasamento si comporta “come deve”, cioè realizza lo sfasamento di π rad: questa frequenza viene “selezionata” dalla rete di sfasamento e l’oscillazione, una volta innescata, prosegue indefinitamente a spese della potenza erogata dal generatore che alimenta il circuito.

Anche senza servirsi di un modello raffinato, è evidente che la fase di oscillazione deve essere preceduta da un transitorio nel quale le condizioni di operazione del sistema “tendono” a divenire oscillanti: lo studio di questo transitorio, al quale si farà cenno nella prossima sezione, non è tra gli obiettivi dell'esercitazione pratica, per cui non entreremo nei dettagli della sua discussione. Però è evidente che esso ha a che fare con la presenza di condensatori e con il tempo che è necessario affinché essi giungano a condizioni stazionarie. Per esempio, supponendo che inizialmente tutti i condensatori siano scarichi, è evidente che un impulso di tensione applicato alla base del transistor passa in parte anche attraverso la rete di sfasamento e da qui a terra. Quindi, in termini molto qualitativi, nella fase iniziale del transitorio solo una frazione dell’impulso viene amplificata dal transistor, per cui v_{out} tende ad avere un livello più alto rispetto a quello che poi si registra quando l'auto-oscillazione ha raggiunto condizioni stazionarie.

VII. REALIZZAZIONE PRATICA CON ARDUINO

Tra i suoi scopi l'esercitazione ha quello di raccogliere dati per farne eventuale uso nell'esercizio sulla FFT. Infatti le oscillazioni prodotte possono essere di qualche interesse in quell'ambito.

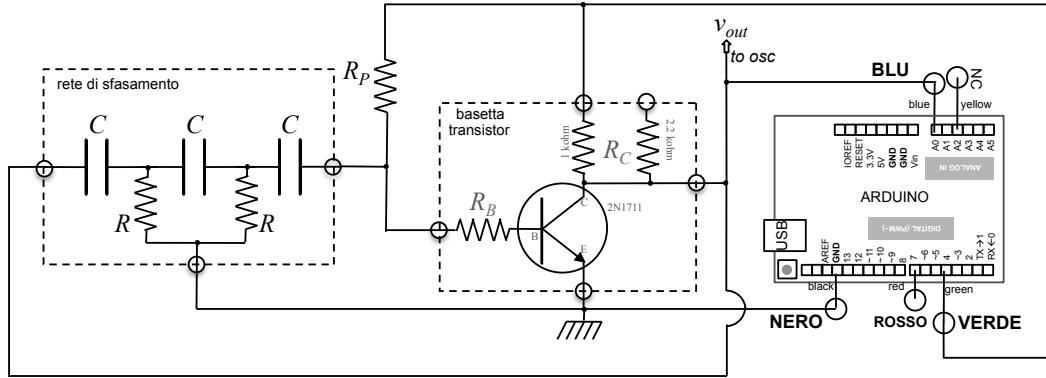


Figura 8. Schema del circuito completo delle connessioni alla scheda Arduino.

Per registrare segnali in funzione del tempo dobbiamo impiegare Arduino. Allo scopo di (i) evitare che il segnale da registrare esca fuori dalla dinamica di Arduino (che digitalizza d.d.p. fino a $V_{max} \sim 5$ V), (ii) favorire l'inesco delle auto-oscillazioni, (iii) permettere una facile sincronizzazione, il circuito viene alimentato da Arduino usando una delle sue porte digitali (pin 5 - boccola verde), che, quando accesa, è attesa trovarsi proprio a V_{max} . Il circuito complessivo è mostrato in Fig. 8. Non esiste alcun motivo fisico per cui il segnale in uscita dall'amplificatore, v_{out} , superi V_{max} , e, analogamente, non esiste alcun motivo per cui esso possa assumere valori negativi: tutto questo garantisce che la porta A0 di Arduino (boccola blu), usata per la digitalizzazione, non riceva mai d.d.p. al di fuori della dinamica consentita. Si consiglia comunque di tenere collegato anche l'oscilloscopio, in modo da visualizzare la forma d'onda prodotta e verificare che l'oscillatore oscilli.

Per rendere possibile la registrazione sincrona di record “lunghi”, il pin 5 viene alimentato solo durante l'esecuzione del programma di acquisizione. Dato che è certamente utile verificare con l'oscilloscopio il funzionamento dell'oscillatore prima di lanciare l'acquisizione, la porta digitale pin 7 (boccola rossa) risulta sempre alimentata a V_{max} . Quindi collegando temporaneamente l’“alimentazione” del circuito a questa boccola (in pratica scambiando fisicamente le due boccole) è possibile verificare la presenza di oscillazioni anche senza lanciare il programma di acquisizione (ma ricordate che lo sketch deve essere già stato caricato su Arduino!).

Visto che in questa esercitazione non siamo interessati alla fase transitoria, il programma di acquisizione (sketch e script di nome **transosc**) fa partire la digitalizzazione passato un certo ritardo (di default 200 ms) dopo l'accensione del pin 5: il tempo $t = 0$ viene fissato al termine di questa fase di attesa. La strategia di acquisizione prevede la registrazione di un certo numero (di default 8) di cicli consecutivi: ogni ciclo parte con un ritardo pari all'estremo temporale del ciclo precedente e i dati vengono incollati tra loro in modo da ottenere un unico file che, di default, è costituito da 2048 coppie di dati, tempo (in μ s) e valore digitalizzato (in digit - in questo esperimento

non c'è alcun bisogno di convertire le unità arbitrarie di digitalizzazione in unità fisiche). La costruzione di record lunghi è vantaggiosa per l'analisi tramite FFT. Se siete interessati ai dettagli, sketch e script, molto simili ad altri precedentemente impiegati, sono riportati in Appendice.

A. Esempi di acquisizione

L'esito dell'esperimento dipende da parecchi fattori. Sicuramente la polarizzazione della giunzione BE, controllata da R_P , è un elemento rilevante: qualora il transistor si trovasse a operare “troppo vicino” al regime di interdizione (R_P “troppo grande”) o di saturazione (R_P “troppo piccola”), l'uscita dell'amplificatore tenderebbe a rimanere a valori costanti, rispettivamente prossimi a V_{max} o a zero (ripassate il funzionamento del transistor a emettitore comune per farvene una ragione!). Il valore di R_P che permette l'auto-oscillazione dipende (anche) dallo specifico transistor impiegato, cioè dal suo guadagno in corrente β_f .

La Fig. 9 mostra un esempio di dati acquisiti nella fase transitoria. Per ottenere questa registrazione ho intenzionalmente modificato lo sketch eliminando il ritardo iniziale di 200 ms e impostando un intervallo di campionamento nominale $\Delta t_{nom} = 60 \mu$ s (voi non siete tenuti ad analizzare questa fase e a fare queste modifiche!). Il grafico mostra chiaramente il transitorio iniziale e la tendenza del segnale ad assumere un andamento oscillatorio al proseguire del tempo. Per questo esempio è stato usato un transistor con $\beta_f \sim 100$ (stima grossolana), $R_C = 1.0$ kohm e $R_P = 0.68$ Mohm (nominali). I dati mostrano chiaramente che, dopo un intervallo di oltre 100 ms, l'ampiezza delle oscillazioni ha raggiunto il suo valore asintotico, testimoniano che il transitorio iniziale si è pressoché esaurito.

La Fig. 10 mostra invece l'auto-oscillazione a regime che si ottiene con lo stesso circuito: in questo caso l'intervallo nominale di campionamento è stato posto pari a $\Delta t_{nom} = 200 \mu$ s e solo una frazione dei dati acquisiti (fino a 100 ms, escluso il ritardo iniziale di default di 200 ms) è stata rappresentata nel pannello superiore. Il pannello

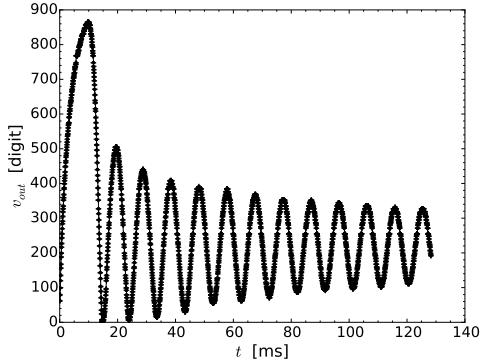


Figura 9. Registrazione del segnale v_{out} durante il transitorio iniziale ($t = 0$ corrisponde nominalmente all'accensione del pin 5 di Arduino) eseguita come descritto nel testo (*voi non siete tenuti ad analizzare questa fase!*). I dati sono rappresentati con punti, dotati delle barre di errore convenzionali (± 1 digit e $\pm 4 \mu\text{s}$), uniti da una linea continua che rappresenta una guida per gli occhi.

inferiore riporta lo *spettro* dello stesso segnale costruito tramite FFT (anche in questo caso si mostra solo una frazione dei dati ottenuti che coprono l'intervallo fino a 1 kHz): per la costruzione della FFT si rimanda alla nota sull'esercizio obbligatorio. In quella nota si chiarisce anche il significato della trasformata di Fourier discreta (numerica) così ottenuta e si mostra come i picchi dello spettro siano rappresentativi delle componenti del segnale alle diverse frequenze. Il pannello mostra chiaramente come ci sia una periodicità dominante a una frequenza attorno a 135 Hz, che dunque è la frequenza di auto-oscillazione del circuito nelle condizioni esaminate. Questa frequenza è in accordo con le previsioni di Eq. 25: nelle condizioni dell'esperimento la corrente di base, misurata con multimetro digitale, risultava $I_B \simeq 1.2 \mu\text{A}$ che corrisponde proprio a valori di $R_{in} = R'$ grossolanamente in accordo con quanto misurato.

La Fig. 11 riporta un altro esempio, realizzato con un transistor diverso e apparentemente dotato di un maggior guadagno in corrente (i parametri del circuito e dell'acquisizione sono specificati in didascalia). I commenti generali sono simili a quelli già riportati per Fig. 9, ma stavolta la frequenza di oscillazione è attorno a 122 Hz, dunque più bassa che nel caso precedente, probabilmente a causa della minore corrente di base (misurata $I_B \simeq 0.9 \mu\text{A}$).

La presenza di diversi picchi nello spettro FFT stimola qualche considerazione. Se il segnale di auto-oscillazione fosse puramente sinusoidale (monocromatico), allora lo spettro mostrerebbe solo la componente fondamentale: invece sono visibili, pur se notevolmente attenuati (notate la scala logaritmica della rappresentazione), dei picchi ad armoniche pari e dispari. Si potrebbe ipotizzare che questo comportamento abbia cause strumentali, legate alla scarsa qualità del campionamento e digitalizzazione effettuate con Arduino. La Fig. 12 mostra il confronto

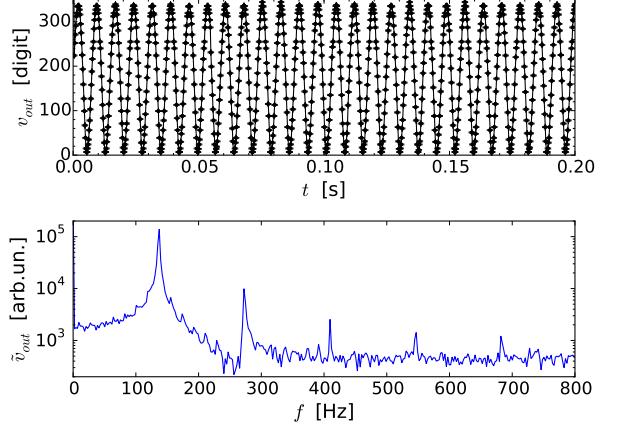


Figura 10. Registrazione del segnale v_{out} che rappresenta l'auto-oscillazione dopo il transitorio iniziale ($t = 0$ è ritardato nominalmente di 200 ms rispetto all'accensione del pin 5 di Arduino) nel pannello superiore e suo spettro \tilde{v}_{out} , ottenuto tramite trasformata di Fourier FFT, nel pannello inferiore. Parametri nominali del circuito e dell'acquisizione: $\beta_f \sim 100$ (stima grossolana), $R_C = 1.0 \text{ kohm}$, $R_P = 0.68 \text{ Mohm}$, $\Delta t_{nom} = 200 \mu\text{s}$. Nel pannello superiore I dati sono rappresentati come in Fig. 9, nel pannello inferiore è usata una linea continua che unisce i punti dello spettro.

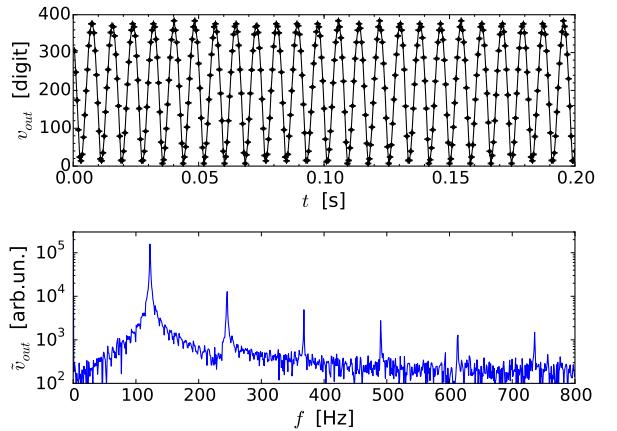


Figura 11. Analogico di Fig. 10, ma con un diverso transistor e diversi parametri nominali del circuito e dell'acquisizione: $\beta_f \sim 200$ (stima grossolana), $R_C = 1.0 \text{ kohm}$, R_P data dal parallelo di 3.3 Mohm e 0.68 Mohm (risultante in $R_P \simeq 0.56 \text{ Mohm}$), $\Delta t_{nom} = 500 \mu\text{s}$.

tra la FFT del pannello inferiore di Fig. 11 e quella di un segnale sinusoidale con caratteristiche (ampiezza, valore medio, frequenza) simili prodotto dal generatore di forme d'onda e acquisito direttamente da Arduino (usando sketch e script di nome *synclong2016*). Si vede chiaramente come i due spettri siano significativamente diversi e come lo spettro del segnale sinusoidale prodotto sia, per la sensibilità della misura, caratterizzato da un solo, stretto, picco alla frequenza impostata sul generatore

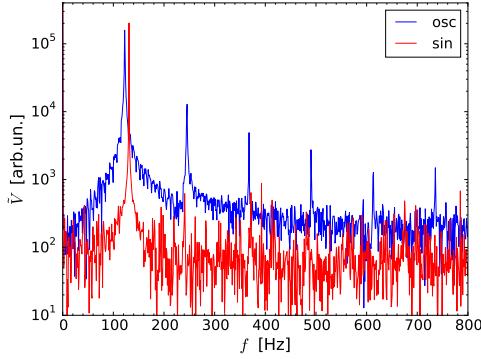


Figura 12. Confronto tra lo spettro FFT del pannello inferiore di Fig. 11, mostrato in blu e indicato in legenda come “osc”, e lo spettro di un segnale sinusoidale a $f \simeq 131$ Hz prodotto dal generatore di forme d’onda e acquisito direttamente da Arduino (mostrato in rosso e indicato in legenda come “sin”).

($f \simeq 131$ Hz).

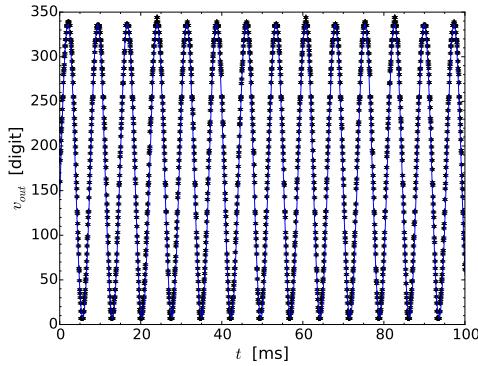


Figura 13. Segnale v_{out} come nel pannello superiore di Fig. 10 con sovrapposto il risultato il best-fit a una funzione sinusoidale (monocromatica), di cui non si riportano i risultati perché non rilevanti in questa sede.

L’analisi FFT suggerisce che la forma d’onda prodotta nell’auto-oscillazione non sia puramente sinusoidale (monocromatica) e permette di farlo con notevole sensibilità. La Fig. 13 mostra lo stesso segnale v_{out} del pannello superiore di Fig. 10 con sovrapposto il best-fit a una funzione sinusoidale (non si riportano i risultati del best-fit perché non rilevanti). Dal punto di vista qualitativo si nota un buon accordo, che dimostra come l’analisi nel dominio dei tempi non sia sufficientemente sensibile per mettere in luce le caratteristiche del segnale di auto-oscillazione, come è invece consentito dalla FFT.

In particolare, la FFT mostra che l’auto-oscillazione è caratterizzata da un intervallo di frequenze, poiché l’oscillazione può avvenire spontaneamente per frequenze che si trovano all’interno di un certo intervallo. Inoltre alla forma dello spettro possono contribuire anche altri dettagli di funzionamento, tra cui il comportamento non

lineare dell’amplificatore a emettitore comune: come già notato in una precedente esercitazione, l’amplificatore ha tendenza a distorcere in modo “asimmetrico” a causa della dipendenza non lineare tra tensione e corrente nella curva caratteristica di base del transistor. Eventuali distorsioni, pur non visibili nell’analisi risolta nel tempo, danno luogo alla presenza delle armoniche superiori nello spettro FFT del segnale v_{out} .

APPENDICE: SKETCH TRANSOSC.INO E SCRIPT TRANSOSC_V1.PY

```

// Blocco definizioni
const unsigned int analogPin=0; // Definisce la porta A0 per la lettura
const unsigned int supplPin = 5; //pin 5 uscita digitale per l'alimentazione dell'oscillatore sincronizzato
const unsigned int supplonPin = 7; //pin 7 uscita digitale per l'alimentazione dell'oscillatore non sincronizzato
int i; // Definisce la variabile intera i (contatore)
int delays; // Definisce la variabile intera delays
int V[256]; // Definisce l'array intero V
long t[256]; // Definisce l'array t
unsigned long StartTime; // Definisce il valore StartTime
unsigned long delayus; // Definisce variabile per acquisizione multipla
unsigned long delayms; // Definisce variabile ausiliaria tempo totale acq
int start=0; // Definisce il valore start (usato come flag)
int j; // Variabile di loop multiacquisizione

// Istruzioni di inizializzazione
void setup()
{
    Serial.begin(9600); // Inizializza la porta seriale a 76800 baud
    Serial.flush(); // Pulisce il buffer della porta seriale
    digitalWrite(supplonPin,HIGH); // Pone supplonPin a livello alto
    //analogReference(INTERNAL); // Sceglie il riferimento V_ref = 1.1 V (nominali)
    bitClear(ADCSRA,ADPS0); // Istruzioni necessarie per velocizzare
    bitClear(ADCSRA,ADPS2); // il rate di acquisizione analogica
}

// Istruzioni del programma
void loop()
{
    delayus=0; // Valori iniziali variabili di ritardo
    delayms=0;
    if (Serial.available() >0) // Controlla se il buffer seriale ha qualcosa
    {
        delays = (Serial.read()-'0')*100; // Legge il byte e lo interpreta come ritardo
        Serial.flush(); // Svuota la seriale
    start=1; // Pone il flag start a uno
    }
    if(!start) return // Se il flag e' start=0 non esegue le operazioni qui di seguito
        // altrimenti le fa partire (quindi aspetta di ricevere l'istruzione
        // di partenza
    delay(1000); // Aspetta 1000 ms per evitare casini
    for (j=0;j<8;j++)
    {
        // delay(1000); // Aspetta 1000 ms per evitare casini
        digitalWrite(supplPin,HIGH);
        delay(200); // Aspetta 200 ms
        StartTime=micros(); // Misura il tempo iniziale con l'orologio interno
        delayMicroseconds(delayus+delays);
        delay(delayms);
        for(i=0;i<256;i++) // Loop di misura
        {
            t[i]=micros()-StartTime; // Legge il timestamp e lo mette in array t
            V[i]=analogRead(analogPin); // Legge analogPin e lo mette in array V
            delayMicroseconds(delays); // Aspetta tot us
        }
        delayms=floor(t[255]/1000);
        delayus=t[255]-delayms*1000;
    }
}

```

```

for(i=0;i<256;i++) // Loop per la scrittura su porta seriale
{
    Serial.print(t[i]); // Scrive t[i]
    Serial.print(" "); // Mette uno spazio
    Serial.println(V[i]); // Scrive V[i] e va a capo
}
digitalWrite(supplPin,LOW);
delay(1000); // Aspetta 1000 ms per evitare casini
}

start=0; // Annulla il flag
Serial.flush(); // Pulisce il buffer della porta seriale (si sa mai)
}

```

```

# Questo script serve per interfacciarsi con Arduino nell'esperienza
# dell'oscillatore oscillatore a reazione con modalita' di acquisizione di record lunghi
# (default 8x256 coppie di dati)

import serial # libreria per gestione porta seriale (USB)
import time # libreria per temporizzazione
import numpy

Directory='../../dati_arduino/' # nome directory dati << DA CAMBIARE SECONDO NECESSITA'
FileName=Directory+'transoscXX.txt' # parte comune nome file << DA CAMBIARE SECONDO NECESSITA'

ard=serial.Serial('/dev/ttyACM0',9600) # apre porta seriale (occhio alla sintassi, dipende
                                         # dal sistema operativo!)
time.sleep(2) # aspetta due secondi per evitare casini
ard.write(b'5') # scrive il carattere per l'intervallo di campionamento
                 # in unita' di 10 us << DA CAMBIARE A SECONDA DEI GUSTI
                 # l'istruzione b indica che e' un byte (carattere ASCII)
time.sleep(1) # aspetta un secondo per evitare casini
outputFile = open(FileName, "w" ) # apre file dati carica per scrittura
runningtime=numpy.zeros(2048)

print ("start")
# loop lettura dati da seriale (256 coppie di dati, ripetuto 8 volte)

for j in range (0,8):
    for i in range (0,256):
        data = ard.readline().decode() # legge il dato e lo decodifica
        if data:
            outputFile.write(data) # scrive i primi 256
            runningtime[i+j*256]=int(data[0:data.find(' ')])
    print ("Part ", j+1, "of 8 done")

outputFile.close() # chiude il file dei dati di carica
ard.close() # chiude la comunicazione seriale con Arduino

# elabora la media e deviazione standard dei delta t

deltat=numpy.zeros(2047)
tsort=numpy.zeros(2048)

```

```
tsort=numpy.sort(runningtime)
for i in range (0,2047):
    deltat[i]=tsort[i+1]-tsort[i]
deltatavg=numpy.average(deltat)
deltatstd=numpy.std(deltat)

print("Delta t average = ",deltatavg," us")
print("Delta t stdev = ",deltatstd," us")

print('end') # scrive sulla console che ha finito
```

Oscillatore smorzato e campionamento con Arduino

francesco.fuso@unipi.it

(Dated: version 12 - FF, 29 marzo 2018)

Questa nota tratta dell'esercitazione pratica su un circuito RLC che si comporta come *oscillatore smorzato*, ponendo particolare enfasi sugli aspetti sperimentali che riguardano l'acquisizione automatizzata del segnale oscillante (e smorzato) con Arduino. Essa copre dunque due aspetti concettualmente diversi, e intende da un lato chiarire e sottolineare dettagli fisici e matematici del circuito e del suo modello, e dall'altro descrivere la, o le, strategie utili per la registrazione automatizzata, tramite Arduino, dei segnali di interesse.

I. CIRCUITO E SUA EQUAZIONE NEL DOMINIO DEL TEMPO

Il circuito considerato è rappresentato in Fig. 1: esso è costituito da un induttore di induttanza L e resistenza interna r collegato a un condensatore C . La parte racchiusa nel box tratteggiato è formata dal generatore di funzioni, configurato in modo da produrre un'onda quadra di frequenza e ampiezza opportune, seguito da un diodo bipolare a giunzione di silicio. Idealmente, e trascurando alcuni dettagli che saranno chiariti in seguito, il diodo si trova in conduzione nella semionda positiva dell'onda quadra, e in interdizione nella semionda negativa. Quindi, semplificando, nella semionda positiva (onda quadra a livello "alto") ai capi del condensatore si trova una d.d.p. approssimativamente pari all'ampiezza dell'onda quadra. Invece nella semionda negativa, quando l'onda quadra è a livello "basso", la maglia di destra, cioè il circuito dell'oscillatore, si trova di fatto isolato (scollegato) rispetto al generatore, visto che il diodo in interdizione è ben approssimato da una resistenza molto alta, idealmente infinita.

In sostanza, allora, la parte racchiusa nel box tratteggiato serve per fornire le necessarie *condizioni iniziali* all'oscillatore armonico smorzato, secondo quanto sarà chiarito nel seguito.

A. Faraday e coefficiente di auto-induzione, o induttanza

Per comprendere il comportamento dell'induttore nel circuito occorre fare riferimento alla cosiddetta legge di Faraday (una delle equazioni di Maxwell, quella del rotore di \vec{E} , scritta in forma integrale). Iniziamo con il notare che l'induttore è, di fatto, un avvolgimento di filo conduttore, ovvero una bobina; chiamiamo $\Phi_S(\vec{B})$ il flusso di campo magnetico *concatenato*, cioè calcolato sulla sezione della bobina stessa. La legge di Faraday stabilisce che

$$\oint \vec{E}^* \cdot d\vec{l} = \varepsilon = -\frac{d\Phi_S(\vec{B})}{dt}. \quad (1)$$

Alla circuitazione del campo elettrico (detto auto-indotto, e per questo indicato con l'asterisco) \vec{E}^* che

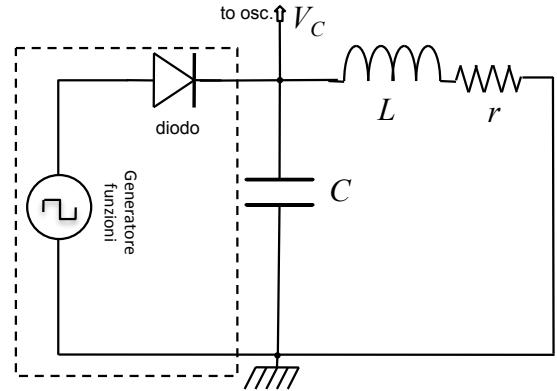


Figura 1. Circuito dell'oscillatore smorzato considerato nel testo.

compare nell'equazione si dà talvolta il nome di *forza elettromotrice*, nome che è abbastanza misleading, se non altro per questioni dimensionali (non si tratta di una forza ma di una differenza di potenziale). Spesso per indicare la forza elettromotrice si usa il simbolo ε , come nell'equazione appena scritta.

Nel caso di nostro interesse, il campo magnetico che compare nella variazione di flusso è quello prodotto dalla corrente che fluisce nell'induttore stesso. In queste condizioni si può porre $\Phi_S(\vec{B}) \equiv LI$, dove L si chiama *induttanza* (o coefficiente di auto-induzione) e I è l'intensità della corrente che attraversa il filo che realizza l'avvolgimento. Supponendo che L sia una caratteristica propria dell'induttore, cioè che dipenda solo dalla sua costruzione (forma, dimensioni, materiale), si ha

$$\varepsilon = -L \frac{dI}{dt}. \quad (2)$$

Il segno meno che compare al secondo membro delle Eqs. 1,2 merita di essere commentato. L'interpretazione qualitativa delle conseguenze del segno meno è che l'induttore reagisce a una variazione di flusso di campo magnetico dovuto alla variazione di corrente "esterna" I in esso inviata dando luogo a un campo magnetico *indotto*, la cui variazione di flusso si oppone a quella del campo prodotto dalla corrente I . Per esempio, supponiamo ci sia una corrente, di verso convenzionalmente positivo, che attraversa l'induttore e immaginiamo che la sua intensità

aumenti nel tempo. Questo dà luogo a un aumento del flusso di campo magnetico. L'induttore *reagisce* in modo da annullare, o tendere ad annullare, questo aumento di flusso. Per farlo, al suo interno si stabilisce una corrente *indotta* che circola in verso opposto rispetto a quella "esterna".

Il passaggio di corrente attraverso l'induttore può essere ottenuto collegando ai suoi capi un generatore di d.d.p. (variabile nel tempo) ΔV_G . In questo modo si forma un circuito a una maglia che è composto dal generatore e dall'induttore. Detta ΔV_L la d.d.p. ai capi del solenoide, deve essere $\Delta V_G = \Delta V_L$. D'altra parte, per l'Eq. 2 sulla maglia, cioè circuitando il campo elettrico sull'intero circuito, si ha anche $0 = \Delta V_G - LdI/dt$. Confrontando le due equazioni si ottiene

$$\Delta V_L = L \frac{dI}{dt}. \quad (3)$$

In questa equazione il segno meno di cui ci siamo occupati prima è scomparso. Il sottile motivo fisico è che tale segno era prima associato alla corrente indotta, mentre nello scrivere le equazioni di un circuito siamo interessati alle correnti "esterne", cioè quelle che inviamo nei vari componenti, le quali determinano la scelta dei segni a seconda del loro verso di scorimento.

L'Eq. 3 può di fatto essere considerata come l'*equazione costitutiva* dell'elemento circuitale che definiamo induttore *ideale*, in cui si trascura l'inevitabile resistenza interna dell'avvolgimento. Essa può essere impiegata per descrivere il comportamento del circuito di Fig. 1 nel dominio del tempo. L'equazione della maglia di destra rappresentata in Fig. 1, *negli istanti di tempo successivi a quello iniziale* t_0 , si può scrivere imponendo che la d.d.p. complessiva sulla maglia sia nulla. Infatti nelle fasi che stiamo considerando il diodo è supposto in interdizione e il generatore di forme d'onda si può considerare scollegato dal circuito. Tenendo conto della caduta di potenziale ai capi della resistenza interna r e della presenza dell'induttore, si ha

$$0 = \frac{Q}{C} - rI - L \frac{dI}{dt}. \quad (4)$$

Ponendo $I = -dQ/dt$, dove il segno negativo tiene conto del fatto che la corrente è dovuta alla carica che *lascia* un'armatura (per esempio, quella superiore rispetto alla figura) del condensatore, e riarrangiando, si ha

$$\frac{d^2Q}{dt^2} + \frac{r}{L} \frac{dQ}{dt} + \omega_0^2 Q = 0, \quad (5)$$

dove si è posto $\omega_0^2 = 1/(LC)$. L'equazione scritta è chiaramente quella di un *oscillatore armonico smorzato*.

Come ben sapete, la soluzione dell'equazione è del tipo $c_1 \exp(\lambda_1 t) + c_2 \exp(\lambda_2 t)$, con $c_{1,2}$ costanti (complesse) che dipendono dalle condizioni iniziali e $\lambda_{1,2}$ soluzioni della cosiddetta equazione caratteristica, che è un'equazione algebrica del secondo ordine:

$$\lambda^2 + \frac{r}{L} \lambda + \omega_0^2 = 0. \quad (6)$$

Si vede facilmente che le soluzioni sono:

$$\lambda_{1,2} = \frac{-\frac{r}{L} \pm \sqrt{\left(\frac{r}{L}\right)^2 - 4\omega_0^2}}{2}. \quad (7)$$

Nel caso di nostro interesse l'oscillatore è, come verificheremo quantitativamente in seguito, *sottosmorzato* e il discriminante dell'equazione di cui sopra è negativo. Posto

$$\omega = \sqrt{\omega_0^2 - 1/\tau^2}, \quad (8)$$

con $\tau = 2L/r$, e $T = 2\pi/\omega$ (detto *pseudo-periodo*), si ha che la soluzione generale di Eq. 5 si può scrivere come $Q(t) = [\exp(-t/\tau)][c_1 \exp(j\omega t) + c_2 \exp(-j\omega t)]$. Poiché la soluzione deve essere reale (quindi $c_1 = c_2^*$), conviene riscriverla come

$$Q(t) = A[\exp(-t/\tau)] \cos(\omega t + \phi) \quad (9)$$

$$A = 2\sqrt{c_1 \cdot c_2} \quad (10)$$

$$\tan \phi = j(c_1 + c_2)/(c_1 - c_2). \quad (11)$$

I coefficienti A (reale) e ϕ (reale) dipendono dalle condizioni iniziali.

È evidente che la soluzione trovata, e, più in generale, lo studio che stiamo conducendo, sono svolti nel *dominio del tempo*. Questo è in effetti necessario, dato che qui ci interessiamo dell'andamento temporale delle grandezze rilevanti (d.d.p., carica, corrente) nell'intervallo che segue all'inizio dell'oscillazione, cioè per $t > t_0$. Vedremo nel seguito che questo circuito è di estremo interesse anche quando il generatore, invece che produrre impulsi, e poi isolarsi per la presenza del diodo, produce onde sinusoidali. In questo caso si è praticamente nelle condizioni di un oscillatore armonico smorzato e *forzato*, e la soluzione a regime può essere determinata in maniera molto efficace usando l'approccio del metodo simbolico, cioè lavorando nel *dominio delle frequenze*.

B. Determinazione di L

Nell'esperienza pratica l'induttore è costituito da una serie di due avvolgimenti concentrici e coassiali tra loro, nominalmente dotati ognuno di 1500 spire (per un totale di 3000 spire se i due avvolgimenti sono in serie). Non è tra gli scopi dell'esperienza, né tra gli obiettivi di questa nota, soffermarsi sulla misura di L , che verrà eseguita in futuro, o andare nei dettagli della determinazione di L a partire dalla geometria degli avvolgimenti. Tuttavia, è opportuno spendere alcune parole su questo aspetto.

In linea di principio, L (così come R per un resistore e C per un condensatore) vorrebbe essere una proprietà del dispositivo realizzato, quindi definita a prescindere dalle condizioni di impiego. Nelle nostre situazioni sperimentali questa affermazione è ragionevole a patto, per esempio, di non spingersi a valori di frequenza superiori a qualche decina di kHz, dove possono avere luogo diversi fenomeni spuri di accoppiamento magnetico.

La determinazione di L a partire dalla geometria della bobina richiede di utilizzare un modello per valutare \vec{B} , e quindi $\Phi_S(\vec{B})$, sulla sezione interna all'avvolgimento. A causa della scarsa simmetria del sistema, questo può essere eseguito, eventualmente, usando un approccio numerico, che è certamente al di fuori degli scopi di questa nota. Qui, invece, ci limitiamo a sottolineare che, per motivi in parte accidentali, è sufficiente utilizzare un modello super-semplificato, e certamente poco realistico, per ottenere un discreto accordo con il valore misurato di L . Il modello in questione è quello del solenoide infinito, per il quale il campo magnetico è presente solo all'interno dell'avvolgimento e ha direzione parallela all'asse della bobina, essendo omogeneo e di modulo $B = \mu_0 NI/\ell$, con N numero di spire e ℓ lunghezza dell'avvolgimento. In queste condizioni si ha $L = \mu_0 N^2 \Sigma / \ell$, con Σ area della sezione della bobina.

Usando dei valori “ragionevoli”, come per esempio $\ell \simeq 6$ cm e $\Sigma \simeq 36$ cm², si ottiene, per $N = 3000$ spire, $L \simeq 0.5$ H, che è in accordo con i tipici valori misurati sulle bobine disponibili in laboratorio quando i due avvolgimenti sono collegati in serie. Per informazione, e, per il momento, senza alcun commento, si riportano anche i valori tipici corrispondenti all'induttanza del singolo avvolgimento interno ed esterno: $L_{int} \simeq 0.1$ H, $L_{ext} \simeq 0.2$ H.

II. ESPERIENZA PRATICA E CONDIZIONI INIZIALI

Nell'esperienza pratica i valori nominali tipici dei vari componenti sono i seguenti: $r \simeq 40$ ohm, *misurata in corrente continua con il multimetro* (torneremo in seguito su questo importante aspetto) e $L \simeq 0.5$ H; inoltre supponiamo $C = 0.1$ μF .

Con questi valori si ha una frequenza angolare *propria* dell'oscillatore $\omega_0 = 1/\sqrt{LC} \approx 4.5 \times 10^3$ rad/s, corrispondente a un periodo *proprio* $T_0 = 2\pi/\omega_0 \approx 1.5$ ms, mentre $1/\tau = r/(2L) \approx 40$ s⁻¹, ovvero $\tau \approx 25$ ms. Di conseguenza l'approssimazione di debole smorzamento è ben verificata e $\omega = \sqrt{\omega_0^2 - 1/\tau^2} \simeq \omega_0$, cioè la pseudo-frequenza angolare dell'oscillatore è molto simile alla sua frequenza propria (e lo pseudo-periodo T è molto simile al periodo proprio T_0). La correzione, infatti, è dell'ordine di poche parti su 10⁴.

Inoltre ricordiamo che nell'esperienza pratica il segnale osservato (vedi Fig. 1) è la d.d.p. V_C presa ai capi del condensatore, il cui andamento temporale previsto è

$$V_C(t) = \frac{Q(t)}{C} = \frac{A}{C} [\exp(-t/\tau)] \cos(\omega t + \phi), \quad (12)$$

dove si suppone che il processo di oscillazione smorzata abbia inizio all'istante $t_0 = 0$.

Come ben sapete, e come già affermato, i parametri A e ϕ che compaiono nella soluzione di Eq. 9 (e di Eq. 12) dipendono dalle condizioni iniziali. Nei tipici esercizi di meccanica la scelta delle condizioni iniziali viene in genere fatta in modo da semplificare l'algebra necessaria per

la determinazione dei parametri incogniti. Per esempio, si sceglie spesso una velocità iniziale nulla o una posizione iniziale che corrisponde a quella di equilibrio. Ovviamennte nel circuito considerato il ruolo della posizione iniziale è quello della carica Q_0 inizialmente (per $t = t_0 = 0$) presente sul condensatore, mentre quello della velocità iniziale è preso dell'intensità di corrente I_0 che scorre inizialmente (per $t = t_0 = 0$) nel circuito.

Dato che

$$I(t) = -\frac{d}{dt} A \exp(-t/\tau) \cos(\omega t + \phi) = \quad (13)$$

$$= A [\exp(-t/\tau)] \left[\frac{\cos(\omega t + \phi)}{\tau} + \omega \sin(\omega t + \phi) \right] \quad (14)$$

deve essere:

$$Q(t = 0) = A \cos \phi = Q_0 \quad (15)$$

$$I(t = 0) = A \left(\frac{\cos \phi}{\tau} + \omega \sin \phi \right) = I_0, \quad (16)$$

ovvero

$$\tan \phi = \frac{1}{\omega} \left(\frac{I_0}{Q_0} - \frac{1}{\tau} \right) \quad (17)$$

$$A = Q_0 \sqrt{1 + \tan^2 \phi}, \quad (18)$$

con $Q_0 = CV_C(t = 0) = CV_{C0}$. Dunque conoscendo Q_0 e I_0 , ovvero V_{C0} (d.d.p. ai capi del condensatore all'istante iniziale) e I_0 , è possibile determinare A e ϕ , o viceversa.

Nel nostro oscillatore la conoscenza di Q_0 e I_0 non è banale, a differenza di quanto succede, in genere, con gli oscillatori meccanici. Possiamo stabilire quali siano i valori di carica e intensità di corrente Q_{in} e I_{in} che si registrano per $t < t_0$: supponendo che il sistema abbia raggiunto condizioni stazionarie nel semiperiodo “alto” dell'onda quadra prodotta dal generatore, chiamando V_A l'ampiezza dell'onda quadra stessa possiamo facilmente porre $Q_{in} = CV_A$ e $I_{in} = V_A/r$, dove abbiamo tenuto conto della circostanza che, in condizioni stazionarie, l'induttore si comporta come una resistenza r . Tuttavia l'istante iniziale è definito all'interno del transiente nel quale l'onda quadra del generatore passa da livello “alto” a livello “basso”. In questo transiente, che nominalmente avviene in un intervallo temporale dell'ordine dei nanosecondi, o decine di nanosecondi (stando al manuale del generatore di funzioni), possono verificarsi fenomeni difficili da modellare, per esempio quelli che coinvolgono il comportamento del diodo.

Consapevoli di questo aspetto, possiamo comunque utilizzare un modello in cui estendiamo per continuità le condizioni iniziali, ponendo $Q_0 = Q_{in}$ e $I_0 = I_{in}$. Come discusso nella prossima sezione, tale affermazione risulta palesemente non verificata sperimentalmente per V_A “grande”, cioè superiore ad alcuni V. In queste condizioni, infatti, il segnale $V_c(t)$ segue un andamento oscillante smorzato solo a partire da un istante $t' > t_0$ (tipicamente l'oscillazione si avvia dopo alcuni ms). A tempi più brevi, infatti, l'oscillazione è rimpiazzata da una veloce discesa

verso valori negativi della d.d.p., che avviene sulla scala delle decine di μs , seguita da una risalita che assomiglia a un esponenziale, con tempi caratteristici di alcuni ms. Questo esponenziale è poi raccordato con l'oscillazione smorzata.

A. Sovratensioni e clipping

Per cercare un significato “fisico” del comportamento che si osserva quando V_A non è “sufficientemente piccolo”, conviene ragionare in termini di energia. Poiché per $t \rightarrow 0^-$ nell’induttore passa la corrente di intensità I_{in} , esso è un “serbatoio di energia” che ha espressione, supponendo I stazionaria, $U_{M,in} = (L/2)I_{in}^2$. Inoltre anche il condensatore è un serbatoio di energia, con espressione $U_{E,in} = (1/(2C))Q_{in}^2$. La possibilità di avere due “tipologie” (due espressioni) di energia non dovrebbe stupirvi troppo: anche in un oscillatore meccanico l’energia complessiva è fatta di due espressioni (potenziale e cinetica, per intenderci).

Se supponiamo, come esempio, $V_A \sim 4$ V, sulla base dei parametri sperimentali prima specificati si ha $U_{M,in} \sim 2.5 \times 10^{-2}$ J e $U_{E,in} \sim 8 \times 10^{-6}$ J. Dunque, nei limiti del nostro modello semplificato in cui trascuriamo il comportamento effettivo nel transiente che definisce t_0 , si ha che il termine di energia magnetico prevale ampiamente su quello elettrostatico. Nell’analogo meccanico, questo indicherebbe un oscillatore dotato inizialmente di energia cinetica molto superiore a quella elastica, cioè, semplificando, un oscillatore le cui condizioni iniziali sono un’elevata velocità e una posizione prossima a quella di equilibrio.

Per $t \geq 0$ entrambi questi serbatoi di energia diventano disponibili per “alimentare” il processo di spostamento delle cariche: a causa della presenza del termine dissipativo (la resistenza), a poco a poco questi serbatoi “si svuotano”. Se, per il momento, trascuriamo la dissipazione, possiamo supporre che l’energia complessiva si conservi. Dunque dopo un quarto di periodo l’energia magnetica sarà nulla (nulla la corrente, ovvero, nell’analogo meccanico, nulla la velocità) e quella elettrostatica varrà, trascurando il valore iniziale, $U_{E,T/4} \sim U_{M,in}$. Poiché $U_E = CV_C^2/2$, da questa uguaglianza si ottiene $V_{C,T/4} \sim \sqrt{2 \times 2.5 \times 10^{-2} \text{J}/0.1\mu\text{F}} \simeq 5 \times 10^2$ V.

Prima di commentare cosa si verifica nell’esperimento, osserviamo che la formazione di una d.d.p. molto elevata ($V_{C,T/4} \gg V_A$) si lega a un fenomeno molto noto in elettrotecnica, quello delle *sovratensioni di apertura*, che possono condurre alla formazione di scariche elettriche quando la corrente che circola in un induttore viene bruscamente interrotta: probabilmente nella vostra esperienza di tutti i giorni c’è l’osservazione di piccole scintille che si creano tra i contatti di un interruttore che controlla un circuito induttivo durante la sua apertura. Queste piccole scintille, dovute alla scarica attraverso l’aria indotta da d.d.p. (e campi elettrici) elevati, possono diventare grandi fonti di rischio nel caso di impianti industriali: spesso,

in questi casi, l’azionamento dell’interruttore è demandato a un servomotore controllato da remoto, in modo che nessun operatore si trovi in prossimità del “coltello” dell’interruttore.

Analogamente, all’atto dell’accensione (chiusura) di un circuito induttivo si può provocare una *sovra corrente di chiusura*, cioè “allo spunto” la richiesta di corrente può essere molto più alta che in condizioni di regime. Anche questa è un’osservazione che potrebbe rientrare nella vostra esperienza quotidiana: probabilmente avrete notato come, all’accensione di un motore elettrico, che è sicuramente un carico induttivo, si possa verificare un temporaneo abbassamento di tensione nella linea di alimentazione (quando avviate il motore di un’automobile, soprattutto di vecchio modello, si abbassa l’intensità delle luci, e qualcosa di simile si ottiene quando si accende l’aspirapolvere in casa), dovuto al temporaneo aumento di richiesta di corrente a un generatore reale (la batteria dell’automobile, la rete di distribuzione elettrica casalinga).

Naturalmente, l’argomento energetico ha un corrispettivo nella legge di Faraday e nell’equazione costitutiva dell’induttore. Per Faraday, la brusca interruzione di corrente spinge l’induttore a reagire creando una corrente indotta che ha lo stesso verso di I_{in} , in modo da minimizzare la variazione di flusso del campo di induzione magnetica: l’induttore “oppone inerzia” alla variazione delle sue condizioni di funzionamento. Per la relazione costitutiva, Eq. 3, alla brusca interruzione di corrente corrisponde una grande d.d.p. ai capi dell’induttore. Tale d.d.p. ha segno negativo, come la derivata temporale, se si attribuisce un segno positivo al verso di circolazione della corrente.

Allora nel breve transiente che definisce l’istante iniziale V_C tende a diventare negativa rispetto alla linea di massa, o terra. Si osserva sperimentalmente che, per fortuna, il suo valore assoluto non giunge alle centinaia di V che abbiamo prima supposto per $V_{C,T/4}$. La spiegazione è molto semplice: l’anodo del diodo resta fisso al valore $-V_A$, e si ha evidentemente $V_{C,T/4} < -V_A$. Dunque il diodo è in conduzione, e quindi non è vero che esso va in interdizione subito dopo che l’onda quadra prodotta dal generatore è passata al suo livello “basso”. Un diodo in conduzione presenta una resistenza interna praticamente trascurabile, per cui in parallelo al condensatore si trova la resistenza del ramo di sinistra del circuito, cioè la resistenza interna del generatore r_G . In sostanza, la presenza del diodo agisce in modo da tosare (*clipping*) il segnale misurato, così che esso non raggiunga mai i valori che, ipoteticamente, potrebbe raggiungere, dato che la d.d.p. ai capi del diodo si stabilizza attorno al valore di soglia ($V_{thr} \sim 0.45 - 0.65$ V).

Volendo, è possibile individuare dei tempi caratteristici che descrivono la rapida discesa di V_C e la sua risalita esponenziale, che avvengono prima che si innescino le oscillazioni smorzate. Nel primo caso, il tempo di caratteristico è quello di carica/scarica del condensatore attraverso la serie $r + r_G$, che vale decine di μs , nel secondo

è quello di “scarica” della corrente I_{in} attraverso la serie dell’induttore e delle resistenze $r + r_G$, che vale alcuni ms.

III. ESPERIMENTO CON ARDUINO

Le prescrizioni per la realizzazione dell’esperimento con Arduino richiedono che V_A sia sufficientemente piccola da impedire che si verifichino i fenomeni sopra descritti. È infatti necessario che sia $|V_C(t)| < 1.1$ V e in queste condizioni la sovratensione di apertura non è mai tale da portare il diodo in conduzione. Quindi la visualizzazione all’oscilloscopio di $V_C(t)$ mostra delle tipiche oscillazioni smorzate che hanno inizio a t_0 (definito, al solito, all’interno del transiente di spegnimento dell’onda quadra). Ovviamente, affinché la visualizzazione sia corretta, occorre scegliere in modo opportuno la modalità di trigger e anche la frequenza del generatore f_G . Essa deve essere tale che, in un suo semiperiodo $T_G/2 = 1/(2f_G)$, abbiano luogo diverse oscillazioni, cioè occorre che $T_G \gg \tau$ (maggiore per un fattore 10 – 20, o quello che preferite).

La visualizzazione all’oscilloscopio permette di misurare lo pseudo-periodo T e anche di determinare τ con l’aiuto di qualche artificio matematico. Tuttavia lo scopo specifico dell’esperienza è quello di *registrare* il segnale $V_C(t)$ con Arduino.

Ci sono almeno tre “problemi” che devono essere subito affrontati affinché questo obiettivo possa essere conseguito.

1. Volendo utilizzare il generatore di forme d’onda per creare in maniera ciclica (periodica, con periodo T_G) le condizioni iniziali, occorre una strategia per *sincronizzare* l’acquisizione di Arduino con l’evento che si vuole analizzare.
2. Poiché Arduino digitalizza solo d.d.p. *positive* (rispetto alla linea di terra), occorre inventare un modo per ottenere un segnale $V'_C(t) > 0$, che sia sempre rappresentativo della oscillazione smorzata.
3. Dato che Arduino impone dei limiti alla massima d.d.p. misurabile (rispetto alla linea di terra, boccola GND), occorre regolare in maniera opportuna l’ampiezza V_A dell’onda quadra prodotta dal generatore (condizione che automaticamente garantisce di poter trascurare i fenomeni legati alla sovratensione di apertura).

A. Sincronizzazione

Per la sincronizzazione, poiché a determinare le condizioni iniziali è il generatore di forme d’onda, possiamo facilmente risolvere il problema utilizzando la lettura *digitale* del segnale di sincronismo presente all’uscita TTL/CMOS OUTPUT del generatore per triggerare la partenza dell’acquisizione.

In particolare, risulta che il livello TTL di questa uscita è “alto” nella fase di interesse per l’acquisizione, cioè quando l’onda quadra prodotta dal generatore si trova nella sua semi-onda negativa e il diodo in interdizione. Lo sketch di Arduino contiene opportuni cicli di attesa che sospendono la partenza dei cicli di misura finché non viene rilevato il passaggio da livello “basso” a livello “alto” di questo segnale di sincronismo, letto sfruttando una porta digitale di Arduino (la 5), configurata come ingresso.

B. Condizionamento del livello

Il modo più semplice e fisicamente “sano” per modificare il livello medio di un segnale, cioè per aggiungervi un offset, o *bias*, V_{bias} costante, consiste nel *sommarlo* alla d.d.p. prodotta da un generatore continuo. Sommare, nel significato che intendiamo dare a questo verbo, significa in sostanza che il segnale $V_C(t)$ deve essere *collegato in serie* a V_{bias} . Naturalmente, dato che il segnale $V_C(t)$ di nostro interesse è riferito alla linea di terra, occorre che il generatore di V_{bias} sia *flottante*, cioè che nessuno dei suoi poli sia collegato a terra.

Questa caratteristica, che si ritrova tipicamente nelle batterie, o pile, è fortunatamente presente anche nell’alimentatore $V_0 \sim 5$ V che tanto spesso impieghiamo in laboratorio. Infatti, nessuna delle sue boccole di uscita è “spontaneamente” collegata alla terra dell’impianto di distribuzione elettrica. Ora, se utilizzassimo direttamente l’alimentatore aggiungeremmo un bias $V_{bias} \simeq 5$ V a $V_C(t)$, che quindi si troverebbe a oscillare (in modo smorzato) attorno a questo valore medio. Usando Arduino, non potremmo fare misure, visto che esso digitalizza segnali di ampiezza massima proprio attorno a 5 V.

La soluzione al problema è rappresentata in Fig. 2: si vede come il generatore di d.d.p. V_0 (o alimentatore, che dir si voglia) sia collegato a un partitore di tensione pre-assemblato, che ha un fattore di partizione $\sim 1/11$ (nominale), permettendo di portare la d.d.p. usata come bias a valori $V_{bias} \simeq 0.5$ V (scarsi).

A questo punto, il valore di V_{bias} implica che, se si vuole che il segnale $V'_C(t) = V_C(t) + V_{bias}$ sia sempre positivo, l’ampiezza *massima* di $V_C(t)$ sia (in valore assoluto) al massimo pari a V_{bias} . Di conseguenza, il massimo range di variazione di $V'_C(t)$ è inferiore a $2V_{bias} \sim 1$ V. Grazie alla possibilità offerta da Arduino di operare con $V_{ref} = 1.1$ V generata internamente, questo range di variazione non implica una perdita significativa di dinamica di digitalizzazione, nel senso che è possibile acquisire i segnali $V'_C(t)$ in maniera sicuramente appropriata.

Alcune osservazioni prima di proseguire:

- dato che V_{bias} è costante, il segnale registrato, $V'_C(t)$ ha esattamente le *stesse* “proprietà” di $V_C(t)$, in particolare le stesse ω , τ (e anche ϕ , fidandosi del metodo di sincronismo enunciato sopra), per cui può essere analizzato in sua vece; la presenza di

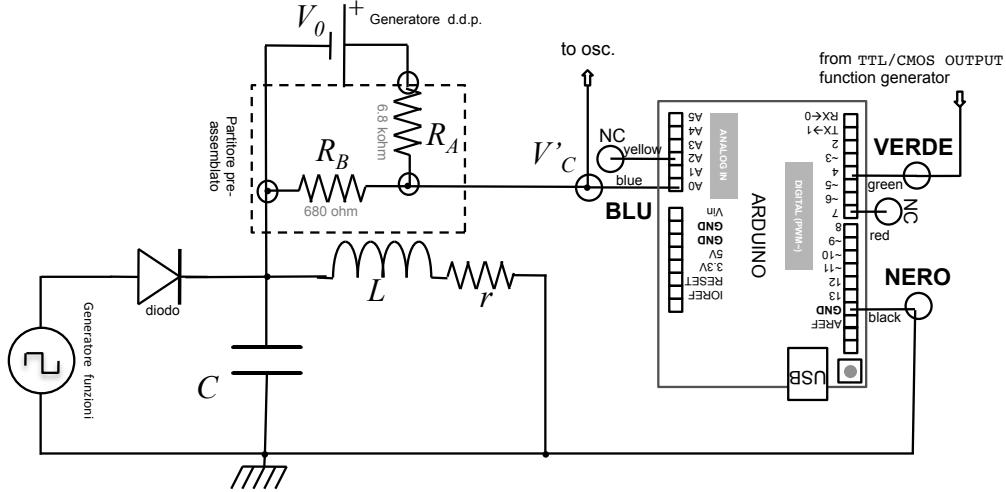


Figura 2. Schema del circuito comprendente l'oscillatore, il generatore della d.d.p. di bias, la scheda Arduino.

V_{bias} può infatti facilmente essere eliminata in fase di analisi, aggiungendo un termine costante alle funzioni di best-fit;

- per le necessità di questa esperienza, non è assolutamente richiesto, anzi, è davvero poco sensato, convertire le unità di digitalizzazione (digit) con cui viene letto il segnale $V'_C(t)$ in unità “fisiche”: non ha quindi senso preoccuparsi di determinare il fattore di conversione relativo e di eseguire l’eventuale propagazione della sua incertezza;
- l’idea molto naïf che la presenza del bias possa essere realizzata aggiungendo un offset all’onda quadra prodotta dal generatore è ovviamente errata, dato che il generatore serve solo per fornire le condizioni iniziali, essendo di fatto scollegato dal circuito durante l’acquisizione dei dati di interesse.

Gli ingressi di Arduino sono *delicati*: dunque *prima* di montare il circuito è *strettamente necessario* controllare con l’oscilloscopio che le condizioni sul bias di cui sopra siano *rigorosamente rispettate*. Il segnale $V'_C(t)$ deve risultare *sempre positivo e di ampiezza massima inferiore a 1.1 V*.

C. Campionamento

Come in ogni occasione in cui abbiamo usato Arduino per campionare dei segnali dipendenti dal tempo, anche per questa esperienza dobbiamo brevemente riflettere sul problema del *sampling rate*. Il sampling rate è, in sostanza, la frequenza con la quale viene eseguito il campionamento (e la successiva digitalizzazione, i due termini sono spesso usati l’uno per l’altro) del segnale analogico in ingresso a un convertitore analogico/digitale. Trascurando per il momento le eventuali latenze, il tempo necessario alla conversione analogico/digitale e i ritardi interni

nell’esecuzione dello sketch, il sampling rate nominale è $S = 1/\Delta t$, con Δt intervallo di campionamento nominale, che anche in questa esperienza può essere impostato nello script di Python usato per il controllo. Ora, anche senza scendere troppo nei dettagli, sappiamo per esperienza che la conversione della tensione analogica nella parola digitale richiede del tempo (stimabile tipicamente in $\Delta t_{conv} \sim 10 - 15 \mu\text{s}$), per cui con l’affermazione appena fatta identifichiamo il sampling rate con l’inverso del ritardo “nominale” tra due campionamenti successivi, piuttosto che con la frequenza effettiva con cui il convertitore analogico/digitale di Arduino opera. Inoltre ogni operazione di digitalizzazione eseguita tramite istruzione software è potenzialmente sottoposta a latenze e ritardi, che possono comportare fluttuazioni nella determinazione degli istanti effettivi di campionamento. Allo scopo di tenere sotto controllo questi ritardi, lo script di Python che gestisce l’esperienza riporta sulla console il valore medio e la standard deviation dell’intervallo effettivo di campionamento, Δt_{eff} , nel campione di dati; in particolare, la standard deviation può essere presa come rappresentativa dell’incertezza sulla determinazione dei tempi (come forse ricordate da precedenti esperienze, essa è tipicamente inferiore a $4 \mu\text{s}$, dunque in prima approssimazione trascurabile rispetto a T e τ).

Il problema della scelta del corretto sampling rate è in genere acuito quando la grandezza da ricostruire è oscillante. Affinché la ricostruzione sia veritiera e utile, occorre che essa sia costituita da un numero “sufficientemente alto” di punti, cioè l’acquisizione deve essere “sufficientemente densa” nel tempo. Nella pratica, una forma sinusoidale ha bisogno di un numero almeno pari a 5–7 punti (meglio una decina) acquisiti all’interno del suo periodo per essere descritta convenientemente. Osservate che, a rigore, questa affermazione è valida solo se vuole costruire un grafico e magari eseguire un best-fit dei dati: infatti, in una qualche forma, esiste un teorema (detto di Nyquist) che stabilisce come in linea di principio, e sapendo

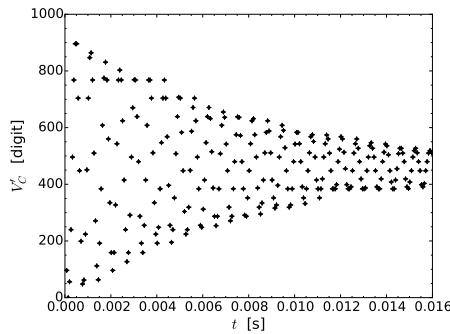


Figura 3. Esempio di acquisizione eseguita con $L = 0.1$ H e $C = 0.1 \mu\text{F}$, $\Delta t = 50 \mu\text{s}$ (tutti dati nominali), scelta per mostrare l'effetto di un sampling rate non pienamente compatibile con i tempi caratteristici del segnale analizzato.

a priori che la forma d'onda è descritta da seni o cosegni, siano sufficienti 2+1 punti per periodo allo scopo di ottenere un buon numero di informazioni rilevanti, per esempio la frequenza di oscillazione determinata tramite trasformata di Fourier (di cui parleremo nel seguito del corso).

L'esigenza di costruire dei grafici rende però stringente la necessità di scegliere un corretto sampling rate. A titolo di esempio, la Fig. 3 mostra un'acquisizione che è “al limite” della rappresentatività. Essa si riferisce all'uso di $L = 0.1$ H (nominale, ottenuta impiegando il solo avvolgimento interno dell'induttore) e $C = 0.1 \mu\text{F}$ (nominali): la pulsazione propria nominale è $\omega_0 = 1 \times 10^4$ rad/s, a cui corrisponde uno pseudo-periodo nominale $T \simeq T_0 \approx 0.6$ ms. L'acquisizione è stata effettuata scegliendo $\Delta t = 50 \mu\text{s}$ nominale, per cui in ogni pseudo-periodo cade circa una decina di punti sperimentali. In effetti non è semplice distinguere nella successione dei punti sperimentali l'andamento sinusoidale atteso. In altre parole, si cominciano a intuire effetti di *sotto-campionamento*.

Nell'esperienza pratica si suggerisce l'uso dell'induttore con i due avvolgimenti in serie, che portano a $L = 0.5$ H, nominale. Per prevenire problemi, il sampling rate della scheda Arduino è stato portato alle sue massime capacità aggiungendo nello sketch le istruzioni che consentono di velocizzare la digitalizzazione, come eseguito, per esempio, nell'esperienza della scarica del condensatore. L'intervallo tra due campionamenti successivi può essere qui aggiustato in unità di $10 \mu\text{s}$ nominali, fino a $\Delta t = 90 \mu\text{s}$ (per andare oltre è necessaria una semplice modifica dello sketch). Generalmente Arduino risponde bene fino a $\Delta t \sim 30 - 40 \mu\text{s}$: al di sotto di tali valori sono possibili *comportamenti erronei*, che in genere sono evidenziati da una crescita anomala della standard deviation riportata sulla console, ben oltre il valore tipico di $2 - 4 \mu\text{s}$.

A causa della limitata capacità di memorizzazione dati di Arduino, anche per questa esperienza la combinazione di sketch e script che vanno sotto il nome di `harm` (rispettivamente `harm.ino` e `harm_v1.py`) consente di

registrare non più di 256 coppie di dati. Questo può porre dei problemi se si vuole osservare lo smorzamento “completo” dell'oscillatore. Per esempio, usando il valore di default $\Delta t = 50 \mu\text{s}$ (nominali), la durata complessiva del record è attorno a 15 ms, quindi minore del tempo di smorzamento τ osservato in alcune condizioni sperimentali.

D. Record lunghi

Come già proposto in una precedente esperienza, la sincronizzazione tra dati e acquisizione e la ripetitività del processo di interesse (che avviene periodicamente con periodo T_G ogni volta che l'onda quadra prodotta dal generatore di forme d'onda passa a livello basso) permettono un'agevole costruzione di record costituiti da un multiplo delle 256 coppie di dati di default. A questo scopo è infatti sufficiente istruire Arduino a compiere un blocco di acquisizioni, scaricarne il risultato sul computer via porta seriale USB, e riavviare automaticamente un nuovo blocco di acquisizioni facendole partire con un ritardo adeguato affinché il primo istante di campionamento sia immediatamente successivo all'ultimo del blocco precedente. Giocando con le istruzioni di Arduino che impongono ritardi, questa operazione è ragionevolmente semplice da implementare (per i dettagli, date un'occhiata agli sketch e script che vanno sotto il nome di `harmlong`). Inoltre, per gli scopi dell'esperienza l'eventuale incertezza addizionale dovuta alla procedura è trascurabile.

Incollando insieme i diversi blocchi così costruiti è possibile ottenere record sufficientemente lunghi per osservare il “completo” smorzamento dell'oscillatore (l'andamento esponenziale decrescente contenuto nella soluzione tende asintoticamente a zero, ma nella misura lo smorzamento “completo” è dovuto al fatto che il segnale misurato scende sotto il livello del rumore, cioè dell'incertezza di misura): questo può dare soddisfazione dal punto di vista estetico. Inoltre, l'uso, per almeno una configurazione sperimentale, di sketch e script progettati per acquisire record lunghi è *necessario* in vista di un futuro impiego con la tecnica FFT. Nel caso, si suggerisce di acquisire record composti dal numero (di default) di 2048 coppie di dati, e di appuntarsi da qualche parte intervallo temporale effettivo di campionamento e sua standard deviation sperimentale, come forniti dalla console.

E. Esempi

Gli esempi riportati nel seguito sono tutti basati sull'acquisizione di 256 coppie di dati (appartengono al “passato”). La Fig. 4 riporta alcuni misure corrispondenti a diverse scelte della capacità C . In tutte le misure:

- è stato impostato l'intervallo di campionamento nominale $\Delta t = 50 \mu\text{s}$; l'intervallo di campionamento effettivo, comprensivo del tempo necessario perché Arduino completi la digitalizzazione, è risultato

mediamente di $62 \mu\text{s}$, come da uscita sulla console dello script di Python, e la standard deviation del campione, pari a $4 \mu\text{s}$, è stata presa come barra di errore per la misura dei tempi.

- Come incertezza per la misura di V'_C è stato usato il valore convenzionale ± 1 digit, che rappresenta una ragionevole sovrastima dell'incertezza stocastica di digitalizzazione, ma una sottostima degli errori di origine prevalentemente sistematica dovuti alle incertezze di campionamento di cui Arduino soffre, specialmente vicino a determinati valori digitalizzati.
- I best-fit sono stati eseguiti rispetto alla funzione modello

$$V'_C(t) = A' \exp(-t/\tau) \cos(\omega t + \phi) + V_{bias}, \quad (19)$$

lasciando parametri liberi A' , τ , ω , ϕ , V_{bias} : è naturalmente possibile ridurre il numero di parametri liberi, per esempio fissando V_{bias} e anche ϕ .

- Nel fit è stata considerata la sola incertezza $\Delta V'_C = \pm 1$ digit; essendo essa (arbitrariamente) attribuita a cause stocastiche, si è usata l'opzione `absolute_sigma = True` nella chiamata alla routine di minimizzazione di Python.
- Nonostante lo sketch preveda la soppressione della prima misura, quella eseguita subito dopo la partenza del trigger dell'acquisizione, le primissime misure iniziali possono essere erronee a causa degli intensi impulsi che girano per il circuito in corrispondenza del fronte d'onda (negativo) dell'onda quadra.
- Di norma generale, occorre un'attenta scelta dei parametri iniziali del fit affinché la routine possa convergere, per cui è fortemente consigliato operare in modalità “esperta”, verificando attentamente la ragionevole congruenza fra funzione di fit calcolata sui parametri iniziali e dati prima di partire (in modalità “apprendista”) con la procedura di minimizzazione.
- Il grafico dei residui normalizzati riporta valori abbastanza spaventosi, che sono probabilmente dovuti sia alla sottostima delle incertezze qui operata, che alla presenza di errori sistematici nella digitalizzazione da parte di Arduino, e, più in generale, nella “sensibilità” delle funzioni armoniche (seno o cosecno) rispetto alla variabile indipendente (il tempo t): è tipico in questi casi ottenere discrepanze non trascurabili tra dati e best-fit, che non dovrebbero terrorizzare nessuno e che potrebbero essere risolte, in modo arbitrario, applicando opportune strategie di rimozione degli outliers.

I risultati dei best-fit (per chiarezza si omettono le covarianze normalizzate, visto il gran numero di parametri liberi impiegato) sono riassunti in Tab. I.

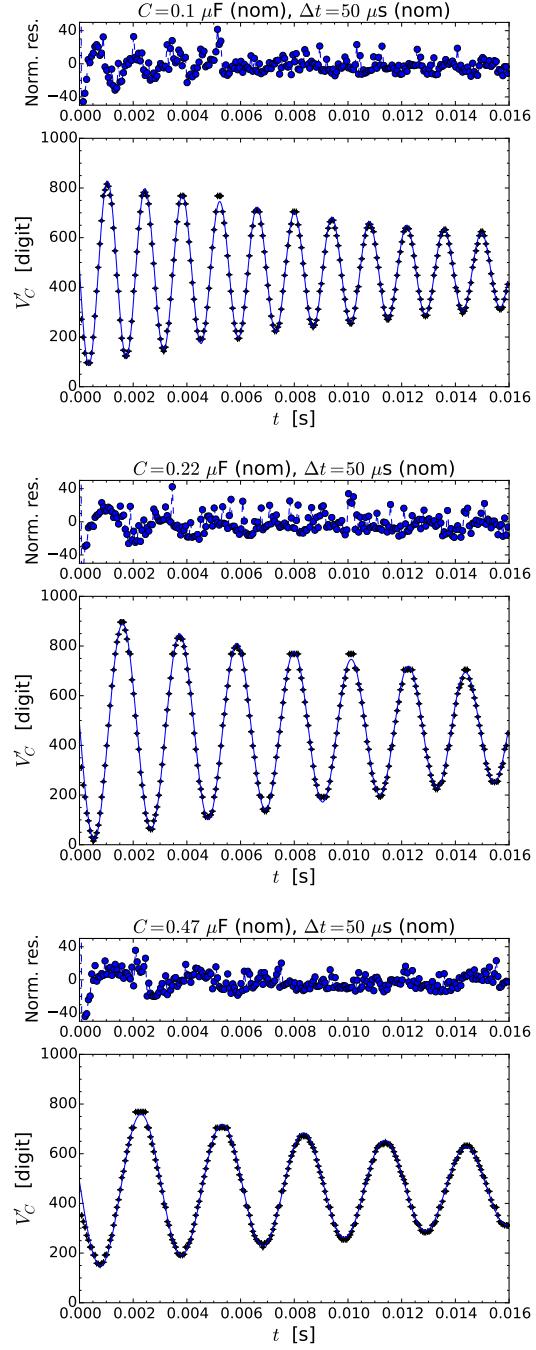


Figura 4. Esempi di misure eseguite con varie scelte di C , come nel titolo dei grafici; in tutti i casi $L = 0.5 \text{ H}$ e $\Delta t = 50 \mu\text{s}$ (tutti dati nominali); i pannelli superiori mostrano l'andamento dei residui normalizzati, le linee continue nei pannelli inferiori sono i best-fit ottenuti come discusso nel testo.

Dal punto di vista generale, i risultati dei best-fit confermano che l'accordo tra dati e best-fit è piuttosto limitato (si vedano i valori del χ^2). Tuttavia, grazie anche all'impiego (arbitrario) dell'opzione `absolute_sigma = True` nella routine di minimizzazione, le incertezze ottenute sui parametri sono relativamente basse (ma si osser-

Tabella I. Risultati dei best-fit dei dati di Fig. 4.

C [μF]	χ^2/ndof	A' [digit]	τ [ms]	ω [krad/s]	ϕ [$\pi/2$ rad]	V_{bias} [digit]
0.1	$3.6 \times 10^5/251$	385.2 ± 0.2	16.34 ± 0.02	4.512 ± 0.008	0.987 ± 0.008	465.7 ± 0.8
0.22	$3.5 \times 10^5/251$	458.8 ± 0.4	20.45 ± 0.06	2.952 ± 0.006	0.987 ± 0.008	466.3 ± 0.8
0.47	$3.1 \times 10^5/251$	329.6 ± 0.3	20.51 ± 0.08	2.076 ± 0.009	0.992 ± 0.006	465.6 ± 0.7

vi che l'uso di numerosi parametri di best-fit, alcuni dei quali fortemente correlati o anti-correlati tra loro, impone cautela nel trarre conclusioni quantitative, poiché la covarianza può influire significativamente nel determinare l'incertezza sulle previsioni). Inoltre i valori di A' , ϕ , e V_{bias} sono in accordo con le aspettative basate su, rispettivamente, l'ampiezza delle oscillazioni (in digit) come osservata nei grafici, lo sfasamento osservato nell'esperienza ($\phi \simeq \pi/2$), il valore di $V_{bias} = (0.498 \pm 0.004)$ V letto con il tester (ricordate che Arduino è operato con $V_{ref} = 1.1$ V nominali, per cui il fattore di conversione è dell'ordine di 1 mV/digit e il valore digitalizzato del bias risulta attorno a 460 – 470).

Come curiosità, avendo eseguito il best-fit, e quindi avendo determinato A e ϕ , possiamo stimare a posteriori i valori di $V_{C0} = Q_0/C$ e I_0 usando le Eqq. 15, 16 [ponendo la debita attenzione al fatto che la grandezza effettivamente misurata non è $Q(t)$, ma $V'(t) = Q(t)/C + V_{bias}$. Trascurando la determinazione (non banale) delle incertezze, si ottiene, per $C = 0.1\mu\text{F}$ nominali (dati di prima riga in Tab. I), $V_{C0} \simeq 8$ mV e $I_0 \simeq 0.2$ mA. Tenendo conto che nelle condizioni dell'esperimento si misurava $V_A \lesssim 1$ V, e quindi $I_{in} \lesssim 25$ mA, si vede come le condizioni iniziali siano sensibilmente diverse dalle aspettative di Sez. II, ma anche come il termine di energia magnetica prevalga ampiamente su quello eletrostatico.

Dal punto di vista più propriamente fisico, i dati di interesse sono quelli di ω e τ , che meritano sicuramente una discussione specifica.

F. ω e τ

Intanto, è evidentemente confermata la circostanza, già più volte anticipata, che l'oscillatore studiato lavora in regime *sotto-smorzato*. La frequenza angolare ω delle pseudo-oscillazioni è attesa, in queste condizioni, essere ben approssimata dalla frequenza angolare propria $\omega_0 = 1/\sqrt{LC}$: le discrepanze sono infatti all'interno dell'incertezza sui parametri ottenuti dai best-fit. Poiché L viene mantenuta costante, il rapporto tra le frequenze angolari trovate è inversamente proporzionale al rapporto tra radici quadrate delle capacità. A causa dell'elevata tolleranza con cui sono note le capacità (la tolleranza è 10% negli esempi riportati), c'è un ovvio accordo con le aspettative. Infatti, per intenderci, $\omega_{0.1\mu\text{F}}/\omega_{0.22\mu\text{F}} = 1.528 \pm 0.007$, che è in accordo con $\sqrt{0.22\mu\text{F}/0.1\mu\text{F}} = 1.5 \pm 10\%$; ana-

logamente $\omega_{0.1\mu\text{F}}/\omega_{0.47\mu\text{F}} = 2.17 \pm 0.01$ è in accordo con $\sqrt{0.47\mu\text{F}/0.1\mu\text{F}} = 2.2 \pm 10\%$.

Sempre con la limitazione dovuta alla scarsa accuratezza nella conoscenza di C , che determina una grossa incertezza nella valutazione, i dati relativi a ω possono essere impiegati per derivare il valore di L : dalle tre misure risulta $L = (0.49 \pm 10\%)$ H, che è pure in accordo con le aspettative (più propriamente, con altre misure, indipendenti, della stessa grandezza, alcune delle quali vedrete nel futuro).

Una conclusione apparentemente inattesa riguarda invece i valori di τ ottenuti dai best-fit relativi alle misure con diverse capacità C . Si vede subito che essi *dipendono* da C , con una tendenza ad aumentare (oltre le barre di errore) con l'aumento di C . Nel nostro modello abbiamo supposto $\tau = 2L/r$, espressione in cui non compare alcuna dipendenza esplicita da C . Inoltre, sulla base dell'andamento di ω da C , possiamo ragionevolmente supporre che L resti costante nel range di parametri sperimentali impiegato, che è anche in accordo con le osservazioni generali sulla definizione di induttanza, o coefficiente di auto-induzione (dovrebbe dipendere solo dalla costruzione dell'induttore).

Fidandoci della valutazione di L appena eseguita, usando il nostro modello possiamo dedurre $r_{0.1\mu\text{F}} = (60 \pm 10\%)$ ohm, $r_{0.22\mu\text{F}} = (48 \pm 10\%)$ ohm, $r_{0.47\mu\text{F}} = (48 \pm 10\%)$ ohm. La misura della resistenza interna dell'induttore fatta con il tester, dunque in *corrente continua*, ha portato, per l'esempio considerato, a $r = (39.2 \pm 0.4)$ ohm, che è minore (oltre le barre di errore) rispetto a tutte le determinazioni di r compiute dai best-fit. Inoltre si osserva un andamento con le condizioni di operazione: in particolare, per il valore di C più piccolo fra quelli impiegati la resistenza r derivata dal best-fit tende ad aumentare significativamente. Notiamo che in queste condizioni anche la frequenza dell'oscillatore tende ad aumentare, per cui possiamo supporre che la discrepanza tra valore di r misurato in continua e quello misurato nelle condizioni dell'esperienza pratica aumenti all'aumentare della frequenza propria di oscillazione del circuito.

G. Dipendenza di r da ω

La resistenza interna dell'induttore, essendo data dalla resistività del (lungo) filo di rame che la costituisce,

dovrebbe anch'essa dipendere solo dalla costruzione (materiale, forma, dimensioni), così come normalmente si afferma per la resistenza di un qualsiasi componente ohmico. Invece la misura della resistenza di un lungo avvolgimento che può operare anche in condizioni di corrente alternata è una tipica situazione in cui tale affermazione non è valida.

Ci sono diversi semplici motivi per supportare questa ipotesi. In primo luogo, all'interno di un qualsiasi filo elettrico percorso da corrente si forma un campo magnetico. I portatori di carica (gli elettroni), che sono sostanzialmente delle particelle cariche in moto prevalente lungo l'asse del filo, risentono della forza di Lorentz, che ha sicuramente anche componenti radiali rispetto al filo. Queste componenti radiali spingono i portatori di carica verso l'asse del filo stesso. Di conseguenza, essi non riempiono più in modo omogeneo la sezione del filo, ovvero, se preferite, la densità di corrente non è più distribuita uniformemente sulla sezione del filo. Se ricordate che, in condizioni di simmetria piana (quella che si applica per un filo cilindrico percorso da corrente omogenea), la resistenza dipende inversamente dalla sezione, l'effetto è quello di aumentare la resistenza effettiva.

Questo fenomeno, che ha molto a che vedere con l'*effetto Hall*, ha conseguenze generalmente poco rilevanti in corrente continua, a meno che le intensità di corrente, e quindi i campi magnetici interni al filo, siano molto elevati. In condizioni alternate, però, la legge di Faraday stabilisce che si formino delle correnti indotte dalla variazione nel tempo del flusso di campo magnetico. All'aumentare della frequenza, cioè, per grandezze armoniche, all'aumentare del valore della variazione temporale, queste correnti possono diventare rilevanti e produrre una d.d.p. che tende a opporsi a quella che fa fluire i portatori di carica nel filo. Di nuovo, l'effetto risultante è quello di ridurre la corrente effettivamente portata dal filo, ovvero, se preferite, di aumentarne la resistenza effettiva.

Questo fenomeno ha, in elettrotecnica, un nome (*effetto pelle*) e un modello descrittivo. Il modello e le sue conseguenze sono riportate in Appendice A, dove si dimostra come, in determinate condizioni, si ottenga una densità di corrente (alternata a frequenza angolare ω) che decresce esponenzialmente muovendosi in direzione radiale dalla periferia all'asse del filo. In altre parole, la corrente scorre prevalentemente (e con distribuzione esponenziale penetrando verso l'interno) sulla "pelle" del filo, cioè in una corona che ha raggio esterno come quello del filo, e raggio interno minore di quello del filo per una certa distanza di penetrazione, detta *profondità di pelle*, che diminuisce all'aumentare della frequenza. Di conseguenza, la sezione del filo effettivamente interessata dalla corrente può essere minore rispetto a quella dell'intero filo, e la resistenza effettiva può aumentare rispetto al valore misurato in continua.

Noteate, en passant, che il motivo per cui i fili elettrici normalmente impiegati (per trasportare anche correnti alternate) sono realizzati con una trecciolina (*trefolo*) di sottili fili di rame ha a che vedere proprio con il desiderio

di limitare le conseguenze dell'effetto pelle. Infatti se il raggio dei sottili fili è minore della profondità di pelle già a frequenze basse, l'eventuale variazione di frequenza della corrente non comporta significative variazioni della resistenza del filo. Impiegando fili sottili, in cui l'anima ha una sezione comunque ridotta, si limitano le variazioni di resistenza effettiva al variare delle frequenze.

Nonostante gli effetti che abbiamo chiamato Hall e pelle siano sicuramente presenti nella nostra esperienza pratica, il diametro del filo (pochi decimi di mm) e le frequenze effettive di lavoro inducono a pensare che essi non giochino un ruolo sostanziale. Infatti, utilizzando stime numeriche, è facile rendersi conto che il confinamento della corrente per effetto Hall e la profondità di pelle conducono a variazioni presumibilmente trascurabili della resistenza r dell'induttore. Occorre allora cercare un'altra motivazione per i risultati ottenuti.

C'è infatti un ulteriore aspetto da considerare quando, come in questa esperienza pratica, si usano degli *avvolgimenti* di filo conduttore. Questo ulteriore aspetto è probabilmente il principale responsabile per l'aumento di r con la frequenza del nostro oscillatore. Infatti è evidente che gli avvolgimenti creano dei campi magnetici anche piuttosto intensi che possono essere "sentiti" dai portatori di carica. Per esempio, nell'ipotesi di solenoidi di lunghezza infinita, gli avvolgimenti più esterni generano dei campi magnetici assiali che insistono nella regione in cui si trovano gli avvolgimenti più interni. Questi campi magnetici generano una forza di Lorentz che ha componenti radiali (rispetto all'avvolgimento). Essa può dunque spingere i portatori di carica verso la periferia dei fili. Anche in questo caso l'effetto è quello di ridurre la sezione del filo interessata dal passaggio dei portatori di carica, ovvero di aumentare la resistenza effettiva. Poiché anche questi campi magnetici dipendono dalla intensità di corrente indotta, a sua volta dipendente dalla frequenza di operazione, la resistenza effettiva può aumentare con la frequenza, come osservato sperimentalmente. Il ruolo che le correnti che scorrono su un filo giocano nel passaggio di corrente in un filo "vicino" dà luogo al cosiddetto *effetto di prossimità* (grazie a Diego!).

Un modo alternativo per rifrassare i fenomeni di cui ci stiamo occupando chiama in causa le cosiddette *correnti parassite*, o *correnti di Foucault*, o, ancora, *eddy currents*. Di queste avremo modo di occuparci in altre esperienze pratiche, dove scopriremo che esse dipendono dalla frequenza e sono sempre accompagnate da un effetto dissipativo. Dato che nel modello del nostro oscillatore la dissipazione è demandata alla resistenza, non stupisce che l'incremento con la frequenza di lavoro degli effetti dissipativi dovuti a tali correnti si rifletta in un aumento del valore della resistenza effettiva "vista" dal circuito.

APPENDICE A: EFFETTO PELLE

Questa Appendice illustra i principali passi che conducono alla soluzione del problema dell'*effetto pelle* e che

permettono di determinare la *profondità di pelle*. In essa vengono usati concetti e strumenti tipici del corso di Fisica Generale 2, che qui vengono dati per noti.

Il problema è posto in questi termini: in un certo riferimento cartesiano, si suppone di avere un vettore densità di corrente \vec{J} diretto lungo l'asse Z e un campo magnetico (di induzione magnetica) oscillante $\vec{B}(t)$ (*sinusoidale* con frequenza angolare ω) diretto lungo l'asse X . Per semplicità di calcolo, si useranno espressioni complesse per il campo magnetico, in modo da poterlo considerare una sorta di fasore (per usare la terminologia a noi cara) e scriverne la derivata temporale come $j\omega\vec{B}$.

L'equazione di Maxwell del rotore del campo elettrico si scrive allora

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}(t)}{\partial t} = -j\omega\vec{B}, \quad (20)$$

e quella per il rotore del campo magnetico \vec{H}

$$\vec{\nabla} \times \vec{H} = \vec{J}. \quad (21)$$

Possiamo modellare il materiale conduttore che sostiene \vec{J} usando la sua resistività, che qui è supposta omogenea e indicata con ρ_c , per cui $\vec{J} = \vec{E}/\rho_c$, dove \vec{E} è il campo elettrico che muove i portatori di carica. Inoltre possiamo supporre di essere nelle condizioni per le quali esiste una relazione lineare tra \vec{H} e \vec{B} , cioè porre $\vec{H} = \vec{B}/\mu$, con $\mu = \mu_0\mu_r$ permeabilità magnetica del mezzo considerato. Di conseguenza le Eqs.20,21 diventano

$$\vec{\nabla} \times \vec{J} = -\frac{j\omega\vec{B}}{\rho_c} \quad (22)$$

$$\vec{\nabla} \times \vec{B} = \mu\vec{J}. \quad (23)$$

Nella geometria cartesiana del problema, i rotori possono essere facilmente esplicitati come derivate spaziali rispetto a una sola coordinata, dando luogo a

$$\frac{\partial J_z(y)}{\partial y} = -\frac{j\omega B_x}{\rho_c} \quad (24)$$

$$-\frac{\partial B_x}{\partial y} = \mu J_z(y). \quad (25)$$

Combinando le due equazioni appena scritte si ottiene la seguente equazione differenziale per $J_z(y)$:

$$\frac{\partial^2 J_z(y)}{\partial y^2} = \frac{j\omega\mu}{\rho_c} J_z(y), \quad (26)$$

La soluzione generale di questa equazione differenziale al secondo ordine è

$$J_z(y) = J_1 \exp(\kappa y) + J_2 \exp(-\kappa y), \quad (27)$$

con

$$\kappa = \sqrt{\frac{j\omega\mu}{\rho_c}} = (1+j)\sqrt{\frac{\omega\mu}{2\rho_c}} = (1+j)k \quad (28)$$

$$k = \sqrt{\frac{2\rho_c}{\omega\mu}}, \quad (29)$$

dove l'ultimo passaggio, quello che conduce a definire la costante k , sfrutta un po' di conoscenze di algebra dei numeri complessi.

Ora, essendo $\kappa > 0$, nella soluzione espressa da Eq. 27 il primo termine della somma non è fisicamente accettabile, poiché conduce a una densità di corrente che aumenta senza limiti (esponenzialmente) con la coordinata y . Definendo $J_0 = J_z(y=0)$ la densità di corrente che si misura nella posizione $y=0$, e tenendo conto dei passaggi di Eq. 28, la soluzione fisicamente accettabile si può scrivere come

$$J_z(y) = J_0 \exp(-y/\delta) \exp(-jy/\delta), \quad (30)$$

dove abbiamo introdotto la nuova grandezza δ , che ha le dimensioni di una lunghezza ed è

$$\delta = \frac{1}{k} = \sqrt{\frac{2\rho_c}{\omega\mu}}. \quad (31)$$

Immaginiamo ora che \vec{J} sia il vettore densità di corrente per una corrente alternata (sinusoidale) che scorre in un filo cilindrico, con il suo asse diretto lungo Z . All'interno del filo, questa corrente provoca un campo magnetico alternato che ha direzione tangenziale. Tracciando una sezione del filo rispetto a un piano che contiene il suo asse geometrico (Z) e la direzione Y ortogonale a questo, il campo magnetico assume la direzione X , come considerato nella dimostrazione condotta. Allora la soluzione espressa in Eq. 30 significa che il modulo della densità di corrente *decade esponenzialmente* mano a mano che dall'esterno del filo si penetra al suo interno muovendosi in direzione *radiale*. La grandezza δ espressa in Eq. 31 rappresenta allora la *lunghezza di penetrazione*, o, meglio, *lunghezza di pelle* della densità di corrente, cioè la distanza alla quale la densità di corrente assume un modulo $1/e$ volte minore di quello misurato sulla superficie del filo stesso.

Alcune doverose precisazioni per concludere:

- il modello impiegato ha delle limitazioni, soprattutto in termini di materiali (considerati omogenei e con resistività indipendente dalla frequenza) e di frequenze (l'effetto pelle ha una descrizione differente per frequenze "alte", cioè in situazioni fortemente non quasi-stazionarie);
- in sostanza, esso descrive in maniera "ragionevole" quanto si osserva in conduttori (fili) di interesse elettrotecnico, cioè di rame, argento, alluminio, e per frequenze di interesse elettrotecnico, cioè fino alle decine/centinaia di MHz.

Tenendo conto di queste precisazioni, l'Eq. 31 fornisce, come esempio, una profondità pelle $\delta \simeq 0.7$ mm per un filo di rame ordinario che porta una corrente alternata (sinusoidale) di frequenza $f = \omega/(2\pi) \simeq 10$ kHz. Nella nostra esperienza, dove si lavora in ogni caso a frequenze più basse (quindi con profondità di pelle maggiore), si fa uso di induttori realizzati con un filo di diametro inferiore a 0.7 mm, per cui l'effetto pelle è atteso avere conseguenze trascurabili.

APPENDICE B: SKETCH HARM.INO

```

// Blocco definizioni
const unsigned int analogPin=0; // Definisce la porta A0 per la lettura
const unsigned int sincPin = 5; //pin 5 ingresso digitale per la sincronizzazione con il generatore
int i; // Definisce la variabile intera i (contatore)
int delays; // Definisce la variabile intera delays
int V[256]; // Definisce l'array intero V
long t[256]; // Definisce l'array t
unsigned long StartTime; // Definisce il valore StartTime
int start=0; // Definisce il valore start (usato come flag)
int sinc; //variabile di sincronizzazione
// Istruzioni di inizializzazione
void setup()
{
    Serial.begin(9600); // Inizializza la porta seriale a 9600 baud
    Serial.flush(); // Pulisce il buffer della porta seriale
    pinMode(sincPin, INPUT); //pin sincPin configurato come ingresso digitale
    analogReference(INTERNAL); // Sceglie il riferimento V_ref = 1.1 V (nominali)
    bitClear(ADCSRA,ADPS0); // Istruzioni necessarie per velocizzare
    bitClear(ADCSRA,ADPS2); // il rate di acquisizione analogica
}
// Istruzioni del programma
void loop()
{
    if (Serial.available() >0) // Controlla se il buffer seriale ha qualcosa
    {
        delays = (Serial.read()-'0')*10; // Legge il byte e lo interpreta come ritardo
        Serial.flush(); // Svuota la seriale
    start=1; // Pone il flag start a uno
    }
    if(!start) return // Se il flag e' start=0 non esegue le operazioni qui di seguito
        // altrimenti le fa partire (quindi aspetta di ricevere l'istruzione
        // di partenza
    delay(2000); // Aspetta 2000 ms per evitare casini
    sinc = digitalRead(sincPin); // legge sincPin
    while (sinc==HIGH) // ciclo di attesa iniziale per sincronizzazione, attende che sincPin vada basso
        {sinc = digitalRead(sincPin);} //legge sincPin
    while (sinc==LOW) // ciclo di attesa iniziale per sincronizzazione, attende che sincPin vada alto
        {sinc = digitalRead(sincPin);} //legge sincPin
    StartTime=micros(); // Misura il tempo iniziale con l'orologio interno
    for(i=0;i<1;i++) // Fa un ciclo di una sola lettura per "scaricare" l'analogPin
    {
        V[i]=analogRead(analogPin);
    }
    for(i=0;i<256;i++) // Loop di misura
    {
        t[i]=micros()-StartTime; // Legge il timestamp e lo mette in array t
        V[i]=analogRead(analogPin); // Legge analogPin e lo mette in array V
        delayMicroseconds(delays); // Aspetta tot us
    }
    for(i=0;i<256;i++) // Loop per la scrittura su porta seriale
    {
        Serial.print(t[i]); // Scrive t[i]
        Serial.print(" ");
        Serial.println(V[i]); // Scrive V[i] e va a capo
    }
}

```

```
}
```

```
start=0; // Annulla il flag
```

```
Serial.flush(); // Pulsice il buffer della porta seriale (si sa mai)
```

Arduino improved e correnti parassite

francesco.fuso@unipi.it

(Dated: version 5 - FF, 31 marzo 2018)

Questa nota ha un doppio scopo: da un lato, essa torna su alcune tematiche tipiche dell'acquisizione dati (con Arduino) e propone alcune possibili soluzioni a problemi comuni in questo ambito, dall'altra tratta dell'esperienza pratica sulle correnti parassite, il cui svolgimento può trarre beneficio dall'applicazione di tali soluzioni.

I. SEGNALE/RUMORE, CAMPIONE DI MISURE, TEMPO DI MISURA

Abbiamo già avuto modo di verificare in diverse esperienze pratiche molte delle problematiche che qui descriviamo. Dunque questa sezione può servire per inquadrare "a posteriori" situazioni, limitazioni e difficoltà probabilmente già note.

È comune che, in fisica e in tutte le altre discipline scientifiche, vengano eseguite misure su sistemi costruiti in laboratorio. Almeno virtualmente, tali sistemi possono spesso essere controllati in modo da essere ripetuti tante volte in maniera automatica. La conseguenza, che differenzia concettualmente le misure "osservative" da quelle "di laboratorio", è che si rende disponibile un *campione di misure*, ovvero che la stessa misura può essere eseguita per un *tempo maggiore*. Questo può comportare un ovvio aumento del rapporto segnale/rumore, fino a rendere intelligibili delle informazioni che altrimenti resterebbero mascherate dal rumore.

Per fare un esempio banale e non del tutto calzante, tutti sappiamo che quando un filo "volante" viene collegato all'ingresso di un oscilloscopio lo schermo mostra un rumore, con componenti anche ad alta frequenza, che indica una variazione della d.d.p. raccolta dal canale di ingresso dell'oscilloscopio. Se lo stesso filo viene collegato a un tester digitale, generalmente non si osserva sul display alcuna variazione della lettura, nonostante il tester abbia una sensibilità certamente non inferiore a quella dell'oscilloscopio. Il motivo è semplice: il tester ha una *banda passante* molto limitata, cioè si comporta da integratore con un tempo di integrazione lungo rispetto al segnale catturato, che quindi viene "mediato a zero" (se alternato, come in genere è il rumore ad alta frequenza). Invece la banda passante elevata dell'oscilloscopio (tipicamente 50 MHz, per i nostri strumenti) permette di eseguire misure su tempi molto più rapidi, in cui alcune componenti ad alta frequenza, quelle al di sotto della banda passante, non fanno in tempo a essere mediate a zero. Rifrasando, se si dà alla misura un tempo sufficientemente lungo per essere eseguita essa permette di annullare, o limitare, gli effetti di alcuni tipi di rumore, in particolare quelli alternati ad alta frequenza. Questo comporta un ovvio aumento del rapporto segnale/rumore, ma, naturalmente, impedisce di seguire la dinamica temporale dei segnali su una scala più rapida del tempo di integrazione dello strumento.

Occorre subito chiarire che esistono quasi sempre dei limiti per l'estensione temporale delle misure: a parte l'ovvia conseguenza di dover aspettare a lungo per ottenere il risultato, acquisire un campione di misure, o aumentare il tempo di misura, significa accrescere la sensibilità nei confronti delle variazioni che avvengono a tempi lunghi. Queste variazioni hanno generalmente a che vedere con *drift* delle condizioni sperimentali, per esempio legati a variazioni di temperatura o della configurazione meccanica, e in genere si sviluppano con tempi caratteristici dell'ordine delle decine o centinaia di secondi. Il loro carattere di inevitabilità, e le molteplici cause fisiche che possono determinarle, hanno spinto a individuarle come una tipologia di rumore, detto qualche volta *flicker noise* o rumore $1/f$, proprio a indicare che il suo contributo diventa sempre più grande a mano a mano che la frequenza della misura, cioè l'inverso del tempo di misura, tende a zero. Naturalmente esistono molte strategie che consentono di mitigare gli effetti di questa tipologia di rumore, ma indubbiamente la sua presenza deve essere tenuta in debita considerazione.

II. ESEMPI DI ACQUISIZIONE IMPROVED CON ARDUINO

Qui di seguito esaminiamo alcuni esempi di strategie di improvement del rapporto segnale/rumore, o comunque del grado di dettaglio delle misure, realizzate con Arduino. L'esperienza di riferimento è quello dell'*oscillatore smorzato rLC*, in cui si richiede di campionare nel tempo il segnale $V_C'(t)$, rappresentativo della d.d.p. presente, istante per istante, tra le armature del condensatore. In questa nota faremo riferimento alle misure svolte impiegando la combinazione di sketch e script denominata *harm* definendole "standard".

L'esperimento possiede naturalmente un carattere *ripetitivo*. Infatti la realizzazione delle condizioni iniziali e la misura si ripetono *periodicamente* nel tempo, grazie all'impiego del generatore di forme d'onda che fornisce un'onda quadra a una certa frequenza f_G . Dunque la misura può essere condotta in tanti intervalli di tempo, di durata massima $T_G/2 = 1/(2f_G)$ (i semiperiodi in cui l'onda quadra è nella sua semionda negativa e il diodo è supposto in interdizione), che si ripetono con un periodo $T_G = 1/f_G$. L'acquisizione tramite Arduino è *sincronizzata* attraverso il segnale TTL prodotto dal generato-

re: quindi sono disponibili tutti gli ingredienti necessari per implementare strategie di acquisizione in grado di sfruttare la ripetitività dell'esperimento.

Un ottimo esempio di tali strategie è rappresentato dalla costruzione di "record lunghi", abilitata dalla coppia sketch e script `harmlong`, basata a sua volta sulla coppia `synclong`, che avete già impiegato. In essa, il carattere ripetitivo e la sincronizzazione fra Arduino e l'esperimento vengono sfruttati per ricostruire il segnale rappresentativo dell'oscillazione smorzata su una durata temporale più lunga, sufficiente per osservare lo smorzamento, mantenendo allo stesso tempo un intervallo di campionamento adeguato per apprezzare compiutamente le oscillazioni.

Qui mostriamo due ulteriori strategie finalizzate a (i) mediare su diversi cicli di acquisizione e (ii) diminuire artificialmente l'intervallo di campionamento.

A. Media e deviazione standard (`harmave`)

La più ovvia, strategia di acquisizione improved nel caso di segnali ripetitivi è quella che prevede di ottenere, dalla ripetizione periodica della misura, i valori della media e della deviazione standard dei dati registrati (la d.d.p. V'_C e il tempo t). I miglioramenti attesi con questo approccio sono sostanzialmente i seguenti:

- il processo di media permette di attenuare gli effetti delle fluttuazioni stocastiche del segnale (gli eventuali errori sistematici, per esempio dovuti a problemi del convertitore analogico/digitale di Arduino, non sono necessariamente influenzati da questo processo);
- la possibilità di calcolare la deviazione standard nel campione di misure consente di stimare in maniera più accurata l'errore (stocastico) delle misure stesse.

La realizzazione pratica di questa strategia è abbastanza immediata. Lo sketch di Arduino, disponibile in rete con il nome `harmave.ino`, è molto simile a quello già impiegato nella acquisizione "standard" (sketch `harm.ino`); lo script di Python (`harmave_v1.py`) contiene invece delle modifiche necessarie a trasferire su arrays i dati letti in ogni ciclo di acquisizione, per poi eseguire il calcolo della media e della deviazione standard e la loro registrazione su un unico file.

Il numero di misure N_{mis} su cui viene eseguita la media può essere impostato nello script (di default le misure sono solo 8), che consente anche di stabilire l'intervallo temporale nominale Δt di campionamento. Lo script produce un file con quattro colonne, che riportano nell'ordine $t, \sigma_t, V'_C, \sigma_{V'_C}$, dove con σ intendiamo la *deviazione standard* sperimentale delle corrispondenti misure sull'intero campione acquisito: i tempi sono misurati in unità di μs , le d.d.p. in unità arbitrarie di digitalizzazione (digit).

Come già affermato, la strategia consente di determinare l'incertezza stocastica delle misure, identificata con

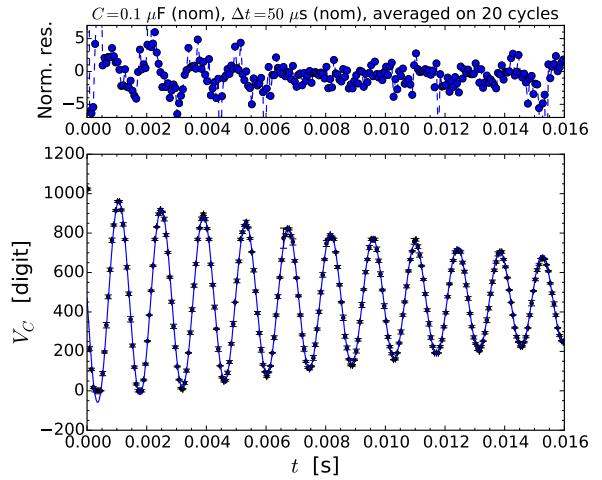


Figura 1. Esempio di dati acquisiti su un oscillatore *rLC* eseguendo la media su $N_{mis} = 20$ cicli, per la scelta di C e Δt (nominali) indicati nel titolo: il pannello superiore mostra il grafico dei residui normalizzati, quello inferiore i dati, correddati delle barre di errore individuate attraverso deviazione standard del campione di misure, e il best-fit. I principali risultati sono $\omega = (4420.73 \pm 0.15)$ rad/s, $\tau = (17.40 \pm 0.02)$ ms, $\chi^2/\text{ndof} = 14991/251$ (non si riportano le covarianze normalizzate per evitare appesantimenti). Il fit è stato eseguito lasciando liberi i parametri $A, \tau, \omega, \phi, V_{bias}$ (si veda la nota riguardante l'oscillatore smorzato *rLC*) e usando l'opzione `absolute_sigma = True`, decisamente consigliata visto il carattere stocastico delle incertezze impiegate.

la deviazione standard sperimentale. Questa procedura può dare origine a un importante problema. Sappiamo già, sulla base delle esperienze svolte in precedenza, che la deviazione standard sperimentale sul valore della d.d.p. digitalizzata da Arduino può essere minore della sensibilità della misura (la singola unità arbitraria di digitalizzazione, o digit). Di conseguenza, specie se il campione di acquisizioni su cui si fa la media è relativamente piccolo, si possono ottenere deviazioni standard *nulle*. Oltre a non avere significato fisico, questa circostanza rende *impossibile* eseguire best-fit con il metodo del minimo χ^2 : infatti la definizione del χ^2 prevede una divisione per la deviazione standard, operazione che in questo caso comporta una divergenza.

Per evitare questo tipo di problemi, lo script di Python contiene un ciclo che analizza il valore di $\sigma_{V'_C}$ (e anche di σ_t) per verificare la presenza di zeri: se esistenti, questi vengono *arbitrariamente* rimpiazzati dal valore medio della deviazione standard sull'intero campione.

La Fig. 1 mostra un esempio dei risultati: in questo caso la media è stata eseguita su 20 misure. Non si discutono qui i risultati del best-fit (riportato con una linea continua nel pannello inferiore), poiché essi non sono rilevanti per la nostra discussione. A livello qualitativo, si riscontra in genere un miglioramento dell'accordo tra dati e best-fit rispetto alle misure "standard", dovuto proprio alla maggiore accuratezza della misura mediata. Tuttavia

gli effetti dei ben noti problemi che Arduino incontra nel campionare segnali variabili nel tempo, specialmente attorno ad alcuni valori specifici del dato digitalizzato, non traggono necessariamente beneficio da questa strategia, come è ovvio tenendo conto della natura *prevalentemente* sistematica di tali problemi.

B. Aumento del sampling rate (`harmint`)

Alcune esperienze pratiche condotte in precedenza hanno indicato che esiste un limite minimo per l'intervallo di campionamento nominale Δt utilizzabile con Arduino. Anche velocizzando il processo tramite “overclock” del microcontroller, come eseguito in questa esperienza, la digitalizzazione diventa instabile, e fornisce risultati potenzialmente erronei se si sceglie un Δt attorno a $20 - 30 \mu s$, o addirittura inferiore.

D'altra parte, la misura dei tempi è un'operazione che può generalmente essere eseguita con grande accuratezza. Infatti, all'interno di Arduino è presente un orologio “al quarzo” che lavora a una frequenza nominale di 16.000 MHz, dove si intende che l'incertezza è sull'ultima cifra significativa. Questo orologio fornisce impulsi di *clock* che distano temporalmente, tra loro, per 62.5 ns (con un'incertezza nominale di pochi ps). Poiché la misura dei tempi viene eseguita *contando* (deterministicamente, cioè in un modo non affetto da rumore analogico) questi impulsi, tale intervallo di tempo costituisce il limite di accuratezza fondamentale delle misure di tempo. Occorre rimarcare, però, che nella realtà l'intervallo di tempo “base” è un multiplo di questo valore (nella nostra configurazione, esso dovrebbe essere o il doppio o il quadruplo, la specifica non è chiaramente ottenibile dai datasheets), a causa delle diverse operazioni che il microcontroller deve eseguire tra un campionamento e il successivo. Inoltre alcune informazioni reperite in rete indicano in $4 \mu s$ l'accuratezza con cui Arduino è in grado di misurare tempi o impostare ritardi, che è ancora minore dei $20 - 30 \mu s$ citati prima.

È possibile realizzare una strategia che consente, grazie al carattere ciclico dell'esperimento, di ottenere intervalli di campionamento effettivi che possono scendere sotto il limite dei $20 - 30 \mu s$. Allo scopo è sufficiente realizzare una acquisizione composta di tanti blocchi successivi, come nel caso esaminato in precedenza. Per ogni blocco, l'intervallo di campionamento è fisso e vale di default, nel nostro caso, $\Delta t = 40 \mu s$ nominali. Quindi si fa in modo che il primo blocco parta, praticamente, in contemporanea con il trigger, mentre i blocchi successivi partono con un ritardo pari al primo a Δt_{int} , il secondo a $2\Delta t_{int}$, e così via. Nello sketch disponibile in rete, Δt_{int} è fisso a $5 \mu s$ nominali e i blocchi acquisiti sono di default 8. È evidente che la procedura non riduce il tempo necessario perché Arduino possa completare la digitalizzazione di un dato, che dipende dalle specifiche costruttive del microcontroller, ma permette di ottenere dati di campionamento su tempi che distano tra di loro, nominalmente, per soli $5 \mu s$, accrescendo artificialmente la sensibilità della misura dei tempi, ovvero la *risoluzione temporale* delle misure.

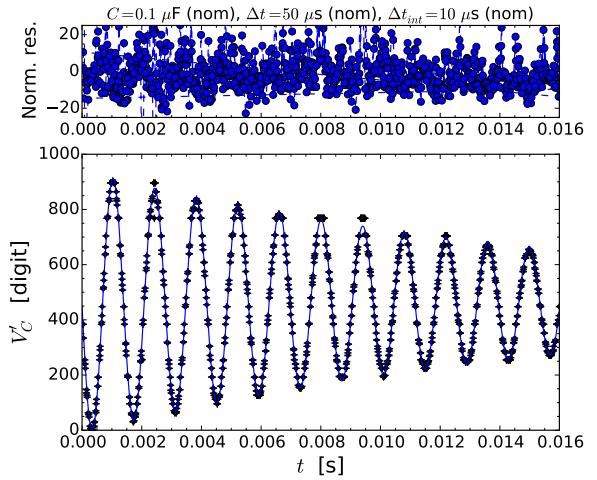


Figura 2. Esempio di dati acquisiti in modalità *interleaved*: per questa presa dati sketch e script sono stati modificati in modo da incollare solo quattro blocchi consecutivi di acquisizione realizzati imponendo un ritardo rispetto al trigger che aumenta con uno step $\Delta t_{int} = 10 \mu s$, in modo da avere una risoluzione temporale effettiva dell'ordine di $10 \mu s$. La scelta di C e Δt nominali è indicata nel titolo: il pannello superiore mostra il grafico dei residui normalizzati, quello inferiore i dati, corredati delle barre di errore (scelte convenzionalmente come $\delta V'_C = \pm 1$ digit e $\delta t = \pm 2.5 \mu s$, valore ottenuto dalla standard deviation di Δt in una acquisizione “standard” effettuata immediatamente prima). I principali risultati sono $\omega = (4505.2 \pm 0.7) \text{ rad/s}$, $\tau = (17.13 \pm 0.02) \text{ ms}$, $\chi^2/\text{ndof} = 4.8 \times 10^5 / 1019$ (non si riportano le covarianze normalizzate per evitare appesantimenti). Il fit è stato eseguito come descritto nella didascalia di Fig. 1.

μs , accrescendo artificialmente la sensibilità della misura dei tempi, ovvero la *risoluzione temporale* delle misure.

Questa modalità di acquisizione si chiama talvolta *interleaved*: una sua raffinata variante fu introdotta già nei primi oscilloscopi digitali commercializzati negli anni '80 dalla ditta svizzera LeCroy (in particolare, una sua variante era presente nel modello 9400, dove consentiva sampling rates fino a 5 GSa/s pur in presenza di un digitalizzatore da soli 100 MSa/s).

Lo sketch costruito per questo scopo è disponibile in rete con il nome `harmint.ino` assieme allo script di controllo `harmint_v1.py`. L'uso pratico dimostra che è possibile ottenere una risoluzione temporale effettiva inferiore a $10 \mu s$, accompagnata, però, da fluttuazioni nella determinazione dei tempi, dovute prevalentemente alle latenze nell'esecuzione dello sketch e dalla accuratezza temporale intrinseca di Arduino. In altre parole, i tempi riportati nel file delle acquisizioni non sono rigorosamente intervallati di $\Delta t_{int} = 5 \mu s$, che è la previsione nominale; inoltre, come importante indicazione pratica, il file prodotto non contiene i dati in forma ordinata crescente nel tempo (i dati non sono sorted secondo i tempi e nel file i blocchi di acquisizione interleaved sono registrati uno di seguito all'altro).

Tabella I. Quadro riassuntivo delle strategie di acquisizione improved disponibili.

Nome sketch	Nome script	Scopo	File prodotto	Colonne del file
<code>harmave.ino</code>	<code>harmave_v1.py</code>	media su N_{mis} misure con calcolo di σ_t e $\sigma_{V'_C}$ da deviazione standard sperimentale (default $N_{mis} = 8$)	256 righe × 4 colonne	$t [\mu\text{s}], \sigma_t [\mu\text{s}], V'_C [\text{digit}], \sigma_{V'_C} [\text{digit}]$
<code>harmlong.ino</code>	<code>harmlong_v1.py</code>	ΔT esteso ($\Delta T_{long} = 8\Delta T$) (default 8 blocchi di acquisizione consecutivi)	256×8 righe × 2 colonne	$t [\mu\text{s}], V'_C [\text{digit}]$
<code>harmint.ino</code>	<code>harmint_v1.py</code>	modalità interleaved con $\Delta t_{int} = 5 \mu\text{s}$, $\Delta t = 40 \mu\text{s}$ (fissati, nominali)	256×8 righe × 2 colonne	$t [\mu\text{s}], V'_C [\text{digit}]$

La Fig. 2 mostra un esempio dei risultati ottenuti operando in modalità interleaved: si vede chiaramente come la densità temporale dei punti acquisiti sia maggiore rispetto all’acquisizione “standard”. L’improvement realizzato in questo caso è quindi nella capacità di acquisire il segnale con una migliore risoluzione temporale, che, almeno in linea di principio, può essere utile per ottenere best-fit più accurati in alcune condizioni sperimentali.

C. Quadro riassuntivo

In sintesi, le denominazioni di sketch e script, gli scopi e le principali caratteristiche delle strategie di acquisizione improved disponibili sono riassunte in Tab. I.

III. BEST-FIT E OUTLIERS

Come già sottolineato, la qualità dei best-fit dei segnali registrati da Arduino, intesa ad esempio come valore del χ^2_{rid} ottenuto, è spesso poco soddisfacente. Probabilmente il motivo principale è nei noti problemi del campionatore di Arduino che danno luogo a errori di natura prevalentemente sistematica, e quindi non sempre riducibili attraverso procedure di media.

In alcuni casi, la presenza di errori sistematici può essere validata con buona confidenza. Per esempio, il primo punto del set di misure, che generalmente ha un valore molto alto, prossimo al massimo valore digitalizzato, può essere interpretato come dovuto alla presenza di *spikes* di corrente, e quindi di d.d.p.. Questi possono essere legati alla brusca variazione delle condizioni di funzionamento dell’oscillatore conseguente al rapido passaggio dalla semionda positiva a quella negativa dell’onda quadra del generatore, che produce un disturbo che si sovrappone alla lettura della d.d.p.. In altri casi, per esempio per i dati acquisiti attorno ai valori 256, 512, 768 digit, la presenza di errori sistematici può essere solo presunta sulla base delle numerose esperienze compiute.

In mancanza di altre forme di controllo, per esempio di “calibrazioni” specifiche della risposta di Arduino nelle condizioni di impiego, l’unica possibilità per evidenziare la presenza di tali errori sistematici è di classificare *arbitrariamente* i dati corrispondenti come *outliers*, ed eventualmente di rimuoverli dal best-fit. Un outlier può essere definito come un dato sperimentale che si discosta per più di un certo valore (arbitrariamente scelto) rispetto alla previsione del best-fit. Ricordiamo qui di seguito alcune avvertenze, a voi probabilmente già note, per la trattazione degli outliers:

- la rimozione degli outliers deve essere motivata;
- essa deve essere sempre *dichiarata*;
- essa deve riguardare solo la procedura di best-fit (il grafico dei dati sperimentali *deve* riportare sempre *tutti* i punti acquisiti);
- gli outliers rimossi dal best-fit *devono* essere indicati nel grafico.

Nello script di Python è possibile inserire istruzioni che permettono di:

1. eseguire un primo best-fit su tutti i dati disponibili;
2. individuare i dati che si discostano in valore assoluto per più di una certa soglia;
3. sovrapporre il grafico degli outliers a quello dei dati sperimentali, facendo in modo che gli uni siano ben distinguibili dagli altri;
4. eseguire un nuovo best-fit sul set di dati senza gli outliers;
5. eventualmente iterare la procedura.

La Fig. 3 mostra un esempio dei risultati. I dati impiegati sono gli stessi di Fig. 1, ma dal best-fit sono stati rimossi i dati lontani per più di tre barre di errore (scelta arbitraria) dal best-fit eseguito considerando l’intero set

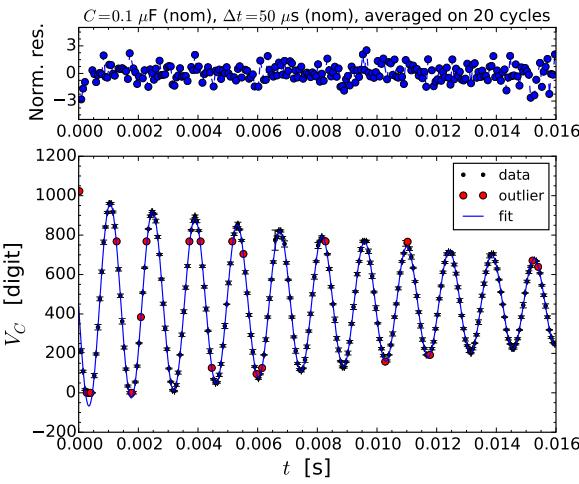


Figura 3. Stessi dati di Fig. 1 trattati con il metodo degli outliers, qui definiti arbitrariamente come i punti distanti per oltre 3 barre di errore dalla previsione del best-fit. Gli outliers sono indicati con pallini rossi nel pannello inferiore, mentre il pannello superiore mostra il grafico dei residui normalizzati. I principali risultati del best-fit sono $\omega = (4420.2 \pm 0.2)$ rad/s, $\tau = (17.55 \pm 0.05)$ ms, $\chi^2/\text{ndof} = 186/230$ (al solito, non si riportano le covarianze normalizzate). Il fit è stato eseguito come descritto nella didascalia di Fig. 1.

di dati. Gli outliers sono rappresentati con pallini rossi. L'effetto della loro rimozione è evidente nel grafico dei residui normalizzati, in cui i “picchi” visibili nel pannello superiore di Fig. 1 sono scomparsi.

Questa procedura porta inevitabilmente a una diminuzione del χ^2 : per i dati dell'esempio si ottiene $\chi^2_{rid} = 0.81$. Supponendo (arbitrariamente) che gli outliers abbiano origine prevalentemente sistematica, e che le incertezze usate nel best-fit siano di origine prevalentemente stocastica (esse sono date dalla deviazione standard sperimentale), si può, per una delle rarissime occasioni che si presentano in questo anno di corso, azzardare arbitrariamente un *test del* χ^2 e affermare, sulla base delle tabelle, che la significatività del best-fit è superiore all’80% (solo in meno del 20% dei casi si otterrebbe un χ^2 minore).

Infine, per poter affermare di aver esplorato ogni possibilità, il best-fit ai dati di Fig. 3 (dunque con la rimozione degli outliers) è stato anche compiuto propagando su V'_C l'incertezza sui tempi, σ_t . I risultati del best-fit non cambiano oltre le rispettive barre di errore, ma il χ^2 si riduce ulteriormente (in questo esempio si ottiene $\chi^2/\text{ndof} = 167/230$).

IV. CORRENTI PARASSITE (E NON SOLO)

Questa sezione della nota si occupa dell'esperienza sulle *correnti parassite* (e non solo). In questa esperienza, oggetti di materiali diversi e di differenti forme e dimensioni vengono inseriti nel core dell'induttore; il segnale

$V'_C(t)$ viene registrato e vengono eseguiti best-fit allo scopo di verificare l'eventuale effetto su ω e, soprattutto, τ della presenza del materiale nel core.

Poiché devono essere eseguiti dei raffronti, è opportuno impiegare sempre la stessa strategia di acquisizione. Negli esempi riportati in seguito la strategia scelta è quella *averaged*: non c'è alcun motivo per questa scelta, e siete invitati, nella vostra esperienza pratica, a decidere autonomamente la strategia che ritenete migliore.

A. Breve background modellistico

La presenza di materiale all'interno del core dell'induttore comporta diversi effetti, che non possono essere modellati con precisione distinguendoli l'uno dall'altro. In una visione molto grossolana, gli effetti attesi possono essere ricondotti a due fenomeni, che qui sono elencati in ordine crescente di rilevanza pratica (per l'esperienza):

1. l'eventuale variazione dell'intensità del campo di induzione magnetica \vec{B} nel materiale, dovuta alle sue “proprietà magnetiche”;
2. la formazione di correnti parassite e i conseguenti effetti in termini di “schermatura” del campo e di dissipazione di potenza.

La corrente che fluisce nell'induttore è responsabile, secondo l'approccio delle equazioni di Maxwell, della creazione di un campo magnetico \vec{H} che, a causa della forma degli avvolgimenti, è particolarmente intenso nel core dell'induttore stesso. In determinate condizioni, che qui supponiamo soddisfatte per i nostri scopi, esiste una relazione lineare tra \vec{B} e \vec{H} :

$$\vec{B} = \mu_0 \mu_r \vec{H}, \quad (1)$$

con μ_0 permeabilità magnetica del vuoto e $\mu_r = \chi_M + 1$ permeabilità relativa del mezzo considerato (χ_M è detta suscettività magnetica).

La definizione di coefficiente di auto-induzione, o induttanza, per il nostro induttore recita $L \equiv \Phi_S(\vec{B})/I$, dove il flusso del campo di induzione magnetica è calcolato sulla sezione S dell'induttore e I è l'intensità della corrente che attraversa l'avvolgimento. A parità di corrente, L viene a dipendere da μ_r : supponendo che l'intero volume del core sia riempito di materiale e supponendo valida l'Eq. 1, L è atteso dipendere linearmente da μ_r .

I materiali impiegati nell'esperienza comprendono (leghe di) alluminio e (leghe di) ferro. Nel primo caso si è in presenza di un mezzo *paramagnetico*. Per i materiali paramagnetici (e anche per quelli diamagnetici), si ha $|\chi_M| \ll 1$ (in particolare, $\chi_{M,Al} \sim 10^{-5}$), per cui $\mu_r \simeq 1$. Di conseguenza, l'eventuale effetto del paramagnetismo (o del diamagnetismo, nel caso di mezzi diamagnetici) è *del tutto trascurabile*.

Per i materiali ferrosi, però, si può avere $\mu_r \gg 1$, anche fino a valori dell'ordine delle centinaia o miglia-

ia. Dunque l'effetto della presenza di un mezzo *ferromagnetico* dovrebbe essere eclatante nelle nostre misure. In una visione molto naïf, L dovrebbe aumentare di ordini di grandezza quando nel core vengono inseriti pezzi di (lega di) ferro, come si fa nell'esperienza pratica. Poiché nell'oscillatore armonico smorzato si ha $\omega = \sqrt{1/(LC) - 1/\tau^2}$, l'incremento di L dovrebbe risultare immediatamente visibile nella sensibile diminuzione di ω .

Anticipiamo da subito che questa sensibile diminuzione (dove per “sensibile” si intende di almeno un ordine di grandezza) *non* si osserva nell'esperienza pratica. Il motivo è collegato alla presenza delle cosiddette *correnti parassite*, o *di Foucault*, o, ancora, *eddy currents*.

La legge di Faraday stabilisce che

$$\oint \vec{E}^* \cdot d\vec{l} = -\frac{d\Phi_S(\vec{B})}{dt}, \quad (2)$$

dove \vec{E}^* è il campo elettrico indotto e la circuitazione è eseguita lunga una linea chiusa che fa da perimetro alla sezione S su cui è calcolata la variazione temporale del flusso del campo di induzione magnetica. Il segno meno presente al secondo membro, ovvero la cosiddetta *legge di Lenz*, significa che le correnti elettriche indotte si muovono in maniera tale da creare un campo magnetico indotto il cui flusso ha una variazione temporale opposta rispetto a quella del campo di induzione magnetica prodotto dalla corrente che circola nell'induttore. In soldoni, e tenendo conto della situazione sperimentale della quale ci occupiamo, le correnti indotte producono un campo magnetico la cui variazione nel tempo è opposta a quella del campo magnetico dovuto alla corrente che viene fatta circolare nell'induttore.

Gli oggetti che, nell'esperienza pratica, infiliamo nel core dell'induttore sono tutti conduttori, chi più e chi meno in funzione del materiale di cui sono fatti e anche dei dettagli costruttivi (se sono pieni, laminati, profilati, tagliati, o altro). Dunque essi sicuramente sostengono, più o meno facilmente a seconda di materiale e geometria, le correnti indotte, che, in questo contesto, prendono proprio il nome di correnti parassite. In linea di principio, esse tendono ad avere la stessa direzione della corrente che scorre nell'induttore. Per Faraday, ovvero Lenz, il loro verso deve però essere opposto a quest'ultima. Allora, all'interno del mezzo, ovvero all'interno della superficie su cui scorrono le correnti parassite, si forma un campo di induzione magnetica indotto che tende a *schermare* il campo di induzione magnetica prodotto dalla corrente che circola nell'induttore, la cui intensità tende quindi a *ridursi* rispetto a quanto si avrebbe in assenza delle correnti parassite. Per questo motivo l'effetto che la presenza del materiale ha su L è meno banale di quanto ci si potrebbe aspettare: in particolare, anche in presenza di materiali ferrosi l'aumento di L non può raggiungere i valori attesi sulla base della conoscenza, o della stima, di μ_r .

Questo effetto di schermatura è sicuramente rilevante. Tuttavia esso non è il principale effetto legato alla

presenza delle correnti parassite. Infatti esse sono collegate principalmente a fenomeni *dissipativi*, dovuti alla circostanza che il materiale in cui scorrono le correnti parassite è un (ordinario) conduttore; pertanto le correnti parassite incontrano una *resistenza*, il che equivale a dire che c'è un fenomeno dissipativo.

Nel nostro oscillatore armonico, la dissipazione è associata al tempo di smorzamento τ la cui riduzione testimonia proprio del verificarsi di dissipazione “aggiuntiva” rispetto a quella dovuta alla resistenza r dell'induttore.

Riassumendo e semplificando per quanto possibile, la presenza del materiale nel core dell'induttore produce, in ordine decrescente di importanza:

- la diminuzione di τ dovuta alla dissipazione prodotta dalle correnti parassite;
- un aumento di ω a causa della diminuzione di τ nella $\omega = \sqrt{1/(LC) - 1/\tau^2}$;
- un'eventuale diminuzione di ω in caso di materiale ferromagnetico e del conseguente aumento di L ;
- un eventuale aumento di ω dovuto all'effetto di schermatura del campo di induzione magnetica \vec{B} nel materiale, provocato dalle correnti parassite.

La relazione tra tutti questi fenomeni e come essi dipendano dal materiale e dalle forme e dimensioni degli oggetti infilati nel core è un argomento difficile da modellare: la misura diretta delle caratteristiche (ω e τ) dell'oscillatore armonico smorzato nelle varie condizioni costituisce quindi un'utile evidenza sperimentale di quanto predetto.

B. Risultati sperimentali

I risultati sperimentali sono stati acquisiti usando la modalità *averaged*, cioè le acquisizioni sono state medicate su $N_{mis} = 20$ misure in modo da ottenere un campione di misure, di cui sono state estratte la media e la deviazione standard, quest'ultima presa come rappresentativa delle barre di errore σ_t e $\sigma_{V'_C}$. I best-fit sono stati eseguiti con la rimozione degli outliers, regolando in questo caso la soglia di esclusione a 5 barre di errore allo scopo di mantenere un numero decente di dati sperimentali nel best-fit. Inoltre nei best-fit è stata eseguita la propagazione dell'incertezza sulla misura dei tempi. In tutti i casi è stato impiegato $C = 0.1 \mu F$, con tolleranza 10%, e si sono usati i due avvolgimenti dell'induttore in serie ($L \simeq 0.5 H$, ovviamente in assenza di materiale nel core). L'intervallo di campionamento nominale è stato regolato a $\Delta t = 50 \mu s$.

La Fig. 4 mostra alcuni esempi delle misure eseguite ponendo diversi oggetti nel core dell'induttore. Al solito, le figure riportano il grafico dei residui normalizzati nel pannello in alto, e i dati sperimentali con il best-fit (gli outliers sono identificati da pallini rossi) nel pannello in basso.

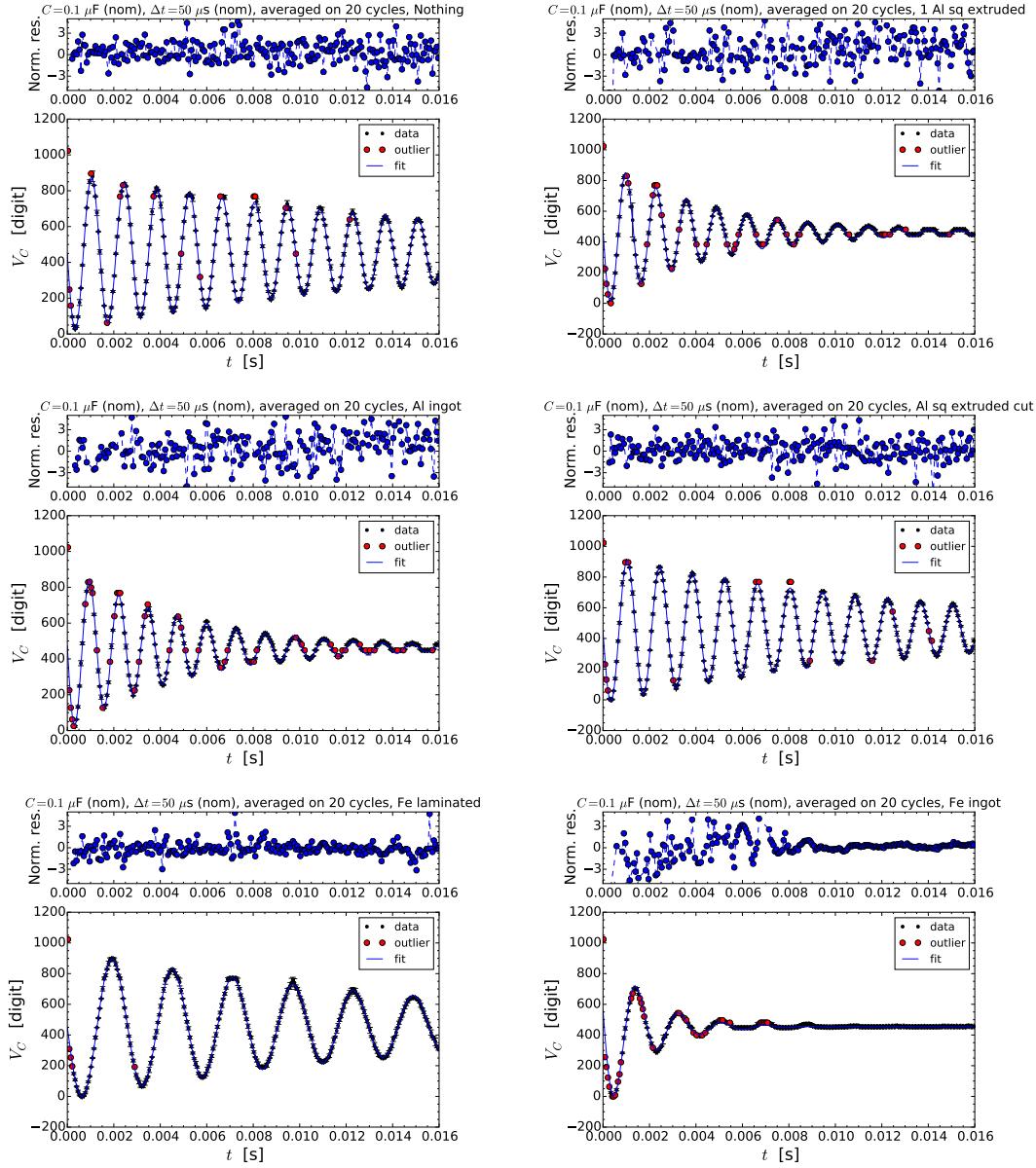


Figura 4. Esempi di misure in presenza di diversi oggetti nel core dell'induttore, come indicato nel titolo dei grafici. La descrizione delle figure e i risultati dei best-fit sono riportati nel testo e in Tab. II.

Si vede chiaramente come i risultati diano luogo a diverse casistiche, che esamineremo dopo aver presentato i principali risultati dei best-fit. Per il momento accontentiamoci di rimarcare alcuni aspetti:

- infilare nel core dell'alluminio non è sufficiente a modificare sensibilmente i risultati: nel caso del profilato segato in longitudinale (indicato con Al sq extruded cut nella figura), l'oscillatore armonico si comporta in maniera simile a quando il core è vuoto (Nothing, in figura);
- tuttavia, nel caso del pieno di alluminio (Al ingot, in figura) e anche del profilato a forma quadrata (1

Al sq extruded, in figura) si osserva una sensibile riduzione del tempo di smorzamento τ ;

- infilando del ferro, poi, si osserva certamente un aumento dello pseudoperiodo di oscillazione, cioè una *diminuzione* della frequenza ω , in particolare nel caso del materiale laminato (Fe laminated, in figura);
- infine, infilando del ferro pieno (Fe ingot, in figura) si ottiene un andamento sensibilmente smorzato.

Possiamo quindi inferire che il comportamento dell'oscillatore in presenza di materiale nel core dell'induttore dipende dal tipo di materiale, ma anche dalla forma e

Tabella II. Principali risultati dei best-fit per diverse scelte degli oggetti infilati nel core dell'induttore. L'ultima colonna riporta il fattore di qualità calcolato usando i parametri ottenuti dal best-fit e propagando le relative incertezze.

Mat	Forma	ω [rad/s]	τ [ms]	χ^2/ndof	L [H] ($\pm 10\%$)	Qf
niente	-	4473.0 ± 0.4	17.50 ± 0.05	372/235	0.50	39.1 ± 0.1
Al	pieno	4977.3 ± 0.7	5.05 ± 0.01	667/201	0.40	12.56 ± 0.04
Al	1 profilato	4816.7 ± 0.8	4.52 ± 0.01	708/211	0.43	10.88 ± 0.06
Al	2 profilati	4824.1 ± 0.6	5.81 ± 0.01	644/236	0.43	14.01 ± 0.03
Al	3 profilati	4813.7 ± 0.5	6.01 ± 0.01	707/233	0.43	14.47 ± 0.03
Al	6 profilati	4809.7 ± 0.6	5.96 ± 0.01	589/235	0.43	14.33 ± 0.04
Al	profilato segato	4482.5 ± 0.4	13.70 ± 0.04	328/235	0.50	30.7 ± 0.1
Al	lamine	4474.1 ± 0.4	16.81 ± 0.04	392/240	0.50	37.7 ± 0.1
Fe	1 lamina	4009.0 ± 0.3	11.54 ± 0.03	364/235	0.62	23.13 ± 0.06
Fe	2 laminae	3819.2 ± 0.3	12.19 ± 0.03	344/238	0.68	23.27 ± 0.06
Fe	3 laminae	3701.5 ± 0.3	12.62 ± 0.04	435/239	0.73	23.35 ± 0.08
Fe	4 laminae	3623.4 ± 0.4	13.39 ± 0.04	399/243	0.76	24.25 ± 0.08
Fe	5 laminae	3551.4 ± 0.3	13.85 ± 0.04	351/246	0.79	24.60 ± 0.07
Fe	laminato	2423.3 ± 0.4	16.13 ± 0.08	181/246	1.7	19.5 ± 0.1
Fe	pieno	3405 ± 2	1.72 ± 0.01	391/215	0.84	2.93 ± 0.02

dalle dimensioni dell'oggetto infilato, il che è in accordo qualitativo con quanto previsto in precedenza.

La Tab. II riporta i principali risultati dei best-fit, incluso il valore di L dedotto dalla $\omega = \sqrt{1/(LC) - 1/\tau^2}$, che si suppone valida per tutti i casi esaminati. Viste le piccole incertezze per i parametri dei best-fit, l'incertezza su L è dominata dalla tolleranza su C . Ovviamente, però, poiché si usa sempre lo stesso condensatore, le eventuali variazioni di L anche minori rispetto all'incertezza sui parametri sono certamente significative (si suppone che il condensatore mantenga le sue caratteristiche da una misura all'altra, cioè si intende la tolleranza sul valore di C come una sorta di errore sistematico, costante di misura in misura).

Si osserva come nel caso dell'alluminio i valori di ω ottenuti siano sempre maggiori del valore registrato in assenza di materiale. Questo suggerisce che l'induttanza non aumenti. In particolare, si osserva come L assuma il suo valore minimo (per il materiale Al) quando viene infilato il lingotto pieno (sezione quadrata di circa 36×36 mm 2). Questa è una conferma dell'effetto di schermatura del campo di induzione magnetica dovuto alle correnti parassite.

Nel caso di impiego dei profilati cavi a sezione quadrata, infilati eventualmente uno nell'altro, si vede come L assuma un valore costante (entro l'ampia incertezza già menzionata). Dunque un solo profilato è sufficiente a sostenere correnti parassite tali da esercitare un forte effetto di schermatura. Infatti infilando ulteriori profilati (di dimensioni trasversali via via decrescenti) non si hanno variazioni di particolare rilievo in L . Ciò è in qualitativo accordo con l'effetto pelle, dato che l'osservazione può essere messa in relazione con lo scorrimento preva-

lentemente superficiale (all'interno di una sottile pelle) delle correnti parassite.

Le correnti parassite sono poi evidentemente collegate a una sensibile diminuzione di τ . È interessante notare che il τ più basso (per il materiale Al), ovvero l'effetto dissipativo comparativamente più rilevante, si ha quando si impiega un singolo profilato. Questo può tentativamente essere interpretato come dovuto alla maggiore resistività del materiale, conseguente al processo di profilatura tramite estrusione. L'aumento della resistività può infatti spiegare l'incremento degli effetti dissipativi. Non è chiaro (ma potete formulare ipotesi) perché infilando ulteriori profilati si osservi un incremento di τ , cioè una diminuzione degli effetti dissipativi.

Per l'oggetto costituito da tante lame di Al unite tra loro con nastro adesivo, l'oscillatore si comporta in maniera molto simile a quando non c'è alcun materiale nel core. Per esercitare i loro effetti, le correnti parassite devono scorrere su un circuito chiuso, cioè devono richiudersi "su se stesse". Per una singola lamina, si può ipotizzare che questo sia più difficile da ottenere. Mettendo tante lame a contatto meccanico, di norma la conduzione elettrica è molto scarsa (l'alluminio ha uno strato di ossido superficiale, per avere contatto elettrico occorre premere con forza in modo da rimuovere meccanicamente questo strato). In altre parole, nel pacco di lame di alluminio le correnti parassite sono fortemente impediti: rispetto alla situazione senza materiale infilato, si osserva un debole aumento degli effetti dissipativi, ma non c'è nessuna evidenza di cambiamenti significativi di ω (e dunque di L).

Simili conclusioni possono essere tratte per il profilato di alluminio segato per lungo: anche qui le superfici a

contatto sono attese presentare una resistenza tale da interrompere, almeno parzialmente, il flusso delle correnti parassite. Questo spiega perché gli effetti riscontrati, sia su τ che su ω , siano molto meno rilevanti che non per il profilato “intero” (non segato).

Infilando del ferro, invece, si osserva generalmente una diminuzione di ω , accompagnata da un aumento di L che va anche ben al di là dell’ampia incertezza associata al suo valore. Usando lamine di ferro, si osservano effetti dissipativi comparativamente molto meno rilevanti che con l’oggetto di ferro pieno. Anche in questo caso si osserva un andamento simile a quello riscontrato per i profilati di alluminio: aggiungendo più lamine (appoggiate l’una sull’altra), si ottiene un aumento di τ , che suggerisce una riduzione degli effetti dissipativi. Allo stesso tempo, si vede come ω diminuisca. La sua diminuzione potrebbe causare una riduzione della forza elettromotrice indotta per effetto Faraday (ricordate che la derivata temporale al secondo membro di Eq. 2 implica che ω vada a moltiplicare), e quindi delle correnti parassite e dei conseguenti effetti dissipativi.

In ogni caso, la presenza delle lamine di ferro provoca un aumento di L . Le lamine impiegate hanno uno spessore nominale di 0.5 mm (e una larghezza paragonabile alla dimensione trasversale del core dell’induttore, che vale circa 36 mm). L’aggiunta di ogni singola lamina provoca quindi il riempimento di circa 1/72 della sezione del core con materiale ferromagnetico. In effetti, L aumenta con il numero di lamine infilate nel core, ma questo aumento non è lineare, probabilmente a causa dell’effetto di schermatura del campo \vec{B} prodotto dalle correnti parassite (seppur deboli vista la geometria laminare del materiale). Inoltre, per il materiale delle lamine ci si aspetta (da altre misure sperimentali) $\mu_r > 10^2$, per cui l’incremento di L dovrebbe essere più brusco di quanto osservato. Di nuovo, la spiegazione della discrepanza può essere ricollegata all’effetto di schermatura di \vec{B} dentro il materiale.

Il valore di L aumenta nel caso dell’oggetto costituito da tante lamine di ferro sovrapposte e tenute assieme con viti e dadi. Questo oggetto è praticamente costituito da circa 25 lamine del tipo prima descritto, per cui quasi la metà della sezione dell’induttore viene riempita di materiale ferromagnetico. In queste condizioni, L aumenta di un fattore 1.5 (abbondante) rispetto all’assenza di materiale nel core. In contemporanea, τ diminuisce di poco. Questo può essere spiegato con l’interruzione delle linee di corrente parassita, operato dall’isolamento elettrico tra le lamine (che sono verniciate).

Infine, usando il pieno di ferro a sezione quadrata (di dimensioni circa $20 \times 20 \text{ mm}^2$) si ha un forte smorzamento, che permette di osservare solo pochi pseudo-periodi di oscillazione. In queste condizioni, τ assume il suo minimo valore in assoluto, risultando circa un ordine di grandezza minore del valore misurato in assenza di materiale. Inoltre, nonostante una gran parte della sezione del core sia riempita di materiale ferromagnetico, il valore di L non è il maggiore in assoluto. Una spiegazione plausibile del-

la prima osservazione è che, in queste condizioni, ci sia una forte dissipazione di potenza causata dalle correnti parassite, legata alla combinazione di elevate correnti e elevata resistività del materiale. Anche la diminuzione dello spessore pelle, che dipende in maniera inversa dalla radice quadrata della permeabilità magnetica, potrebbe avere un ruolo nell’accrescere la dissipazione, dato che correnti parassite confinate in un piccolo spessore di pelle potrebbero incontrare maggiore resistenza che non correnti che penetrano di più all’interno del materiale. Inoltre, è evidente il fenomeno di schermatura del campo di induzione magnetica nel core, che contrasta l’aumento di L dovuto alla permeabilità magnetica. Infine, a causa del legame tra ω e ω_0 , la determinazione di L dipende anche dal valore di τ : in particolare, a parità di ω la diminuzione di τ comporta una diminuzione di L . Poiché nell’osservazione sperimentale si ha che ω diminuisce rispetto alla situazione imperturbata meno di quanto diminuisca τ , è proprio l’effetto dovuto a τ che, dal punto di vista matematico, prevale.

V. FATTORE DI MERITO O QUALITÀ (Q-FACTOR)

La “bontà” di un oscillatore armonico può essere valutata attraverso una specifica quantità adimensionale, il *fattore di qualità*, o *fattore di merito*, o, ancora, *Q-factor*, che indicheremo proprio come Q_f .

Una buona e generale *definizione* di fattore di qualità è la seguente:

$$Q_f = 2\pi \frac{E_{\text{stored}}}{E_{\text{lost per cycle}}} , \quad (3)$$

dove il termine 2π a moltiplicare ha un’origine legata all’uso delle frequenze angolari, e qualche volta viene omesso. Le due grandezze che compaiono nella frazione, rispettivamente al numeratore e al denominatore, si riferiscono all’energia immagazzinata nell’oscillatore e a quella persa (dissipata) per ogni ciclo di oscillazione.

Come ben sappiamo, l’energia in un oscillatore *RLC* è rappresentata da due contributi di diversa tipologia: il contributo elettrostatico, localizzato nel condensatore e proporzionale al quadrato della carica da esso accumulata, e quello magnetostatico, localizzato nell’induttore e proporzionale al quadrato dell’intensità di corrente che lo attraversa.

Poiché carica e intensità di corrente sono legate tra loro dall’operatore derivata temporale, e visto l’andamento sinusoidale nel tempo di queste grandezze, ci saranno sicuramente (tanti) istanti di tempo in cui la carica è nulla e la corrente ha la sua massima intensità I_{\max} (in modulo, qui dei segni possiamo disinteressarci). In questi instanti l’*energia immagazzinata nell’oscillatore* è fatta dalla sola componente magnetostatica; come sapete (e come torneremo in futuro a dimostrare), l’espressione di questa energia, che quindi rappresenta *tutta* l’energia

immagazzinata nell'oscillatore, è

$$E_{\text{stored}} = \frac{L}{2} I_{\text{max}}^2. \quad (4)$$

La presenza del componente dissipativo (la resistenza effettivamente presente nel circuito, ovvero quella “apparente” dovuta alle correnti parassite e alla loro dissipazione), che qui indichiamo come R , implica perdita di potenza per effetto Joule. In un oscillatore sotto-smorzato, in cui si trascura l'energia dissipata in un ciclo, a potenza media dissipata per effetto Joule è

$$\langle P_{\text{Joule}} \rangle = \frac{R I_{\text{max}}^2}{2}. \quad (5)$$

Basandoci sulla potenza dissipata mediamente per ciclo, possiamo a questo punto esprimere l'*energia persa per ciclo* come

$$E_{\text{lostpercycle}} = \langle P_{\text{Joule}} \rangle T = \frac{R' I_{\text{max}}^2}{2f}, \quad (6)$$

dove $T = 1/f$ rappresenta il periodo, o, meglio, lo pseudo-periodo, dell'oscillazione armonica sotto-smorzata.

Si ottiene quindi:

$$Qf = 2\pi \frac{LI_{\text{max}}^2/2}{RI_{\text{max}}^2/2} f = 2\pi \frac{L}{R} f = \omega \frac{L}{R}. \quad (7)$$

Ricordando che per un oscillatore RLC (qui con R si indicano “tutte” le resistenze presenti, vere e dovute alle correnti parassite) è $\tau = 2L/R$, si può anche scrivere

$$Qf = \frac{\omega\tau}{2}. \quad (8)$$

In pratica, e a parte coefficienti moltiplicativi, il fattore di qualità è proporzionale al rapporto tra tempo di smorzamento e periodo: un oscillatore che presenta poca dissipazione, che cioè ha un tempo di smorzamento molto lungo rispetto al periodo, ha un Qf superiore a un oscillatore che, invece, si smorza in fretta rispetto al periodo. Un oscillatore che ha un fattore di qualità molto basso (in genere si considera $Qf \leq 1/2$) non è più sotto-smorzato.

Nel mondo, come avremo modo di sottolineare ancora, esistono oscillatori armonici che possono avere Qf molto elevati, per esempio negli orologi (convenzionali e atomici) e nei laser. Il nostro oscillatore RLC , invece, è di qualità generalmente scarsa. A titolo di esempio, i valori di Qf registrati nelle varie configurazioni esplorati sono riportati in Tab. II. Si noti che, grossolanamente, il fattore di qualità corrisponde al numero di cicli di oscillazione “chiaramente” visibili (cioè ben distinti dal rumore) che si osservano nelle acquisizioni con Arduino, ovvero all'oscilloscopio. Quindi una stima grossolana di Qf può essere compiuta in maniera diretta, contando i picchi che si osservano in ogni ciclo di oscillazione.

Risonanze

francesco.fuso@unipi.it

(Dated: version 8.1 (piccola correzione di Eq. 36) - FF, 25 marzo 2017)

Questa breve nota tratta alcuni aspetti (soprattutto di calcolo) relativi all'esercitazione sull'oscillatore RLC forzato e smorzato e propone qualche spunto di riflessione ulteriore sulla tematica della risonanza.

I. INTRODUZIONE

Il fenomeno della risonanza è probabilmente tra quelli che hanno la più ampia applicazione in fisica e anche in tante altre discipline (ehm, si direbbe nel linguaggio comune che la risonanza ha la più ampia risonanza...). Inutile fare un elenco delle situazioni in cui la risonanza occupa un posto di rilievo ed è anche inutile ricordare gli effetti eclatanti della risonanza (per esempio, i ponti che crollano che campeggiano nei libri delle scuole medie, esempio di risonanza in realtà piuttosto opinabile). L'aspetto più importante del fenomeno è che, in risonanza, si stabilisce un efficace trasferimento di potenza da una sorgente (la forzante) a un sistema (qui modellato come un oscillatore armonico smorzato).

Per fare riferimento specifico a quanto svolto nell'esperienza pratica, il circuito che si comporta da oscillatore smorzato e forzato, mostrato in Fig. 1, comprende due resistori che tengono conto della resistenza interna dell'induttore, r , e di una resistenza esterna, R . Inoltre lo schema mostra anche la presenza di un ulteriore elemento resistivo, la resistenza interna del generatore r_G (non si mostra invece la resistenza interna dello strumento di misura, l'oscilloscopio, che riteniamo produca effetti trascurabili). Sono anche indicati i segnali (differenze di potenziale) misurate nell'esperienza, V_{in} e V_{out} .

Detta $V_{0G} \cos(\omega t) = V_G(t)$ la differenza di potenziale applicata alla maglia, l'equazione del circuito nel dominio del tempo è:

$$\frac{d^2Q(t)}{dt^2} + \frac{R'}{L} \frac{dQ(t)}{dt} + \frac{1}{LC} Q(t) = \frac{V_{0G}}{L} \cos(\omega t), \quad (1)$$

con $R' = r + R + r_G$. La soluzione di questa equazione può sicuramente essere condotta nel dominio del tempo con i metodi "ordinari" a voi già noti: si ottiene che esiste una soluzione oscillante a frequenza ω .

Se ci restringiamo a considerare la soluzione a regime, quella che si stabilisce dopo un transiente iniziale (di durata paragonabile a $\tau = 2L/R'$, un tempo che, nelle nostre osservazioni sperimentali, è generalmente breve rispetto a quello di misura), allora possiamo fare ottimo uso del metodo simbolico, che consiste, in sostanza, nell'indicare come grandezze complesse (oscillanti) tensioni e correnti. L'equazione del circuito in questo contesto si scrive

$$V_{\omega G} = Z_{tot} I_\omega \quad (2)$$

dove $V_{\omega G}$ e I_ω sono i fasori che rappresentano rispettivamente il segnale prodotto dal generatore, supposto ideale,

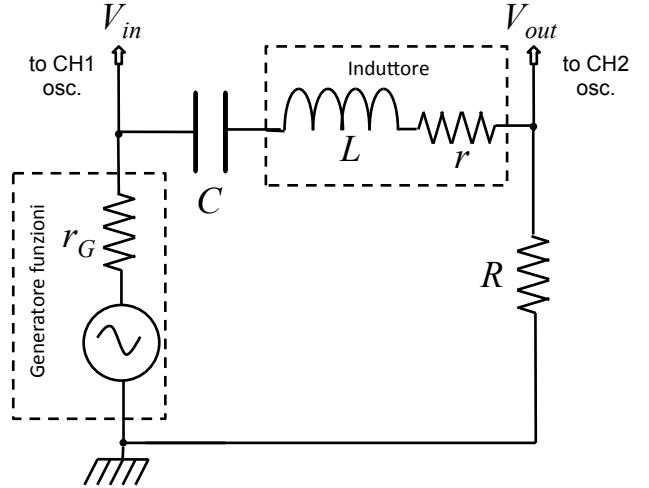


Figura 1. Circuito dell'oscillatore risonante (serie) considerato nel testo.

e la corrente che circola nella maglia è

$$Z_{tot} = R' + \frac{1}{j\omega C} + j\omega L = \frac{j\omega R' C + 1 - \omega^2 LC}{j\omega C} \quad (3)$$

è l'impedenza totale del circuito, dato dalla serie dei componenti che costituiscono la maglia.

In sostanza, quindi

$$V_{\omega G} = \frac{j\omega R' C + 1 - \omega^2 LC}{j\omega C} I_\omega. \quad (4)$$

Il segnale indicato con V_{out} in figura è, supponendo l'impedenza interna dell'oscilloscopio molto più grande delle altre impedenze in gioco, come in genere si verifica, la caduta di tensione ai capi della resistenza R . Dunque, tornando alla notazione simbolica e facendo un po' di pulizia, si ha

$$V_{\omega out} = \frac{j\omega RC}{j\omega R' C + 1 - (\omega/\omega_0)^2} V_{\omega G}, \quad (5)$$

con $\omega_0^2 = 1/LC$. Possiamo quindi determinare una funzione di trasferimento complessa $T_G(\omega)$ tale che $V_{\omega out} = T_G(\omega) V_{\omega G}$:

$$T_G(\omega) = \frac{j\omega RC}{j\omega R' C + 1 - (\omega/\omega_0)^2}. \quad (6)$$

Si vede subito che, per $\omega = 1/\sqrt{LC} = \omega_0$, un pezzo del denominatore si annulla, cosa che, almeno intuitivamente, ha a che fare con la risonanza.

Nella pratica, generalmente si preferisce definire la funzione di trasferimento come quella che lega i segnali osservati sperimentalmente, cioè V_{out} e V_{in} ; quest'ultimo segnale differisce da V_G per la caduta di potenziale sulla resistenza interna del generatore r_G , cioè si ha, per i fasori corrispondenti:

$$V_{\omega in} = V_{\omega G} - r_G I_\omega = V_{\omega G} \left(1 - r_G \frac{j\omega C}{j\omega R'C + 1 - (\omega/\omega_0)^2} \right), \quad (7)$$

dove nell'ultimo passaggio abbiamo usato l'espressione di I_ω trovata prima.

Invertendo l'equazione e usando un minimo di algebra si trova

$$V_{\omega G} = \frac{j\omega R'C + 1 - (\omega/\omega_0)^2}{j\omega(R+r)C + 1 - (\omega/\omega_0)^2} V_{\omega in}, \quad (8)$$

che, sostituita nell'Eq. 5, conduce a

$$V_{\omega out} = \frac{j\omega RC}{j\omega(R+r)C + 1 - (\omega/\omega_0)^2} V_{\omega in}. \quad (9)$$

La funzione di trasferimento complessa $T(\omega)$ tale che $V_{\omega out} = T(\omega)V_{\omega in}$ è allora

$$T(\omega) = \frac{j\omega RC}{j\omega(R+r)C + 1 - (\omega/\omega_0)^2}. \quad (10)$$

È evidente che essa approssima e viene approssimata da quella espressa in Eq. 6 se $R' \sim (R+r)$, cioè, come atteso, se la resistenza interna del generatore è trascurabile, ovvero $r_G \ll (R+r)$.

A. Risonanza nella serie e nel parallelo condensatore/induttore

Prima di proseguire con quel poco di matematica che serve a caratterizzare la risonanza, ricordiamo qui da dove questo fenomeno ha origine nella tipologia di circuiti di cui ci stiamo occupando, con lo scopo principale di puntualizzare tutti gli aspetti più semplici e immediati da capire. A questo scopo, facciamo riferimento a una situazione ideale, in cui immaginiamo che sia possibile avere un induttore in cui la resistenza interna dovuta all'avvolgimento del filo sia completamente trascurabile. In altre parole, immaginiamo di avere un componente che presenta un'impedenza solo *reattiva*, con un'induttanza di valore L . Supponiamo che questo componente sia *in serie* a un condensatore C . L'impedenza complessiva della serie è $Z_{tot} = 1/(j\omega C) + j\omega L = (1/(j\omega C))(1 - \omega^2 LC)$: per un certo valore della frequenza (angolare), cioè a *risonanza*, ovvero per $\omega = 1/\sqrt{LC} = \omega_0$, l'impedenza totale Z_{tot} si annulla.

Se ragionate in termini di rappresentazione dell'impedenza sul piano complesso, la situazione è semplice da

descrivere: l'impedenza del condensatore è un vettore disposto lungo l'asse immaginario e orientato verso il basso (verso negativo), quella dell'induttore è un vettore di analoga direzione ma verso opposto. I moduli di questi vettori, cioè delle impedenze considerate, dipendono dalla frequenza: a una certa frequenza, cioè a risonanza, essi sono uguali e la somma dei due vettori, che è l'impedenza risultante, si annulla.

In una maniera ancora più limpida e convincente, che fa sempre uso della rappresentazione sul piano complesso, notiamo che nel collegamento in serie è il fasore corrente I_ω a essere in comune ai due elementi (induttore e condensatore). Il fasore $V_{\omega C}$ che rappresenta la d.d.p. ai capi del condensatore risulta anticipato di $\pi/2$ rispetto a I_ω , e di lunghezza (modulo) dipendente in modo inversamente proporzionale alla frequenza. Il fasore $V_{\omega L}$ è invece ritardato di $\pi/2$ rispetto a I_ω , e la sua lunghezza è direttamente proporzionale alla frequenza: i due fasori che rappresentano le d.d.p. sono quindi mutamente sfasati di π . Per una data frequenza, quella di risonanza, la somma delle due d.d.p., $V_{\omega serie} = V_{\omega C} + V_{\omega L}$, rappresentata da due fasori antiparalleli tra loro, si annulla, cioè la serie LC si comporta come un cortocircuito, dunque con impedenza nulla.

Vediamo brevemente cosa succede se i due componenti considerati vengono collegati in parallelo. Stavolta l'impedenza totale è $Z_{tot} = (j\omega C + 1/(j\omega L))^{-1} = (j\omega L)/(1 - \omega^2 LC)$: in questo caso alla risonanza l'impedenza tende a infinito, cioè non c'è corrente che passa attraverso il parallelo dei due componenti. Infatti, usando lo stesso approccio di prima, stavolta è la d.d.p. V_ω a essere in comune ai due componenti montati in parallelo. È facile rendersi conto che, impiegando una simbologia facilmente comprensibile, in questo caso $I_{\omega C}$ è posticipata di $\pi/2$ rispetto a V_ω , mentre $I_{\omega L}$ è ritardata di $\pi/2$. Di nuovo i moduli dei fasori che rappresentano la corrente che passa per il condensatore e per l'induttore hanno un modulo dipendente dalla frequenza e risultano sfasati reciprocamente di π . A una data frequenza, quella di risonanza, le due correnti sono opposte, e si ha che $I_{\omega parallelo} = I_{\omega C} + I_{\omega L}$ si annulla. Il parallelo allora si comporta come un circuito aperto, ovvero la corrente fluisce nel parallelo di induttore e condensatore e non ne esce fuori.

Ovviamente se aggiungiamo un elemento resistivo, fosse anche solo la resistenza interna r dell'induttore, le affermazioni appena fatte si modificano in maniera rilevante. Per esempio è facile rendersi conto che nel circuito serie anche a risonanza c'è un'impedenza non nulla (la r , che è in serie a tutto il resto) e che nel circuito in parallelo anche a risonanza l'impedenza è finita (la r che è in serie alla L , ovvero in parallelo a C). Di conseguenza, nel caso reale ci si aspetta che gli effetti della risonanza siano meno eclatanti. Infatti, come ben sapete dallo studio di altre situazioni fisiche, considerare una "dissipazione" (e la resistenza è un componente che "dissipa" per effetto Joule), cioè un attrito, conduce sempre ad attenuare gli effetti della risonanza.

B. Risposta in frequenza

La funzione di trasferimento $T(\omega)$ di Eq. 10, che è complessa, fornisce informazioni sia sull'attenuazione, o guadagno, del circuito, che sullo sfasamento del segnale in uscita rispetto a quello in ingresso attraverso rispettivamente il suo modulo, che indichiamo con $A(\omega)$, e il rapporto tra parte immaginaria e parte reale, che indichiamo con $\tan(\Delta\phi)$.

Razionalizzando la funzione espressa nell'Eq. 10 si ha

$$T(\omega) = \frac{j\omega RC[(1 - (\omega/\omega_0)^2) - j\omega(R + r)C]}{(1 - (\omega/\omega_0)^2)^2 + (\omega(R + r)C)^2}. \quad (11)$$

Lo sfasamento tra V_{out} e V_{in} dipende dalla frequenza secondo la:

$$\tan(\Delta\phi) = \frac{\text{Im}\{T(\omega)\}}{\text{Re}\{T(\omega)\}} = \frac{1 - (\omega/\omega_0)^2}{\omega(R + r)C}. \quad (12)$$

Si vede subito un aspetto molto molto significativo: *lo sfasamento si annulla a risonanza* e cambia di segno passando per la risonanza stessa ($\phi \rightarrow \pm\pi/2$ rispettivamente per $\omega \rightarrow 0$ o $\omega \rightarrow \infty$). Ciò ha una conseguenza estremamente interessante in termini *sperimentali*: a risonanza, lo sfasamento è nullo. Lo sfasamento può essere osservato facilmente (e con ottima sensibilità) usando diversi strumenti, il più semplice dei quali consiste nel visualizzare i segnali V_{in} e V_{out} all'oscilloscopio in modalità X-Y: sullo schermo appare in genere una ellisse che, a risonanza, degenera in un segmento *inclinato*. Questo consente di individuare rapidamente la frequenza di risonanza di un oscillatore armonico forzato.

Veniamo ora alla determinazione di $A(\omega)$, cioè troviamo quella che spesso si chiama la *curva di risonanza*, o lo *spettro di risonanza*, o, ancora, la *forma di riga* del nostro oscillatore. Si ha:

$$A(\omega) = |T(\omega)| = \frac{\omega RC}{\sqrt{(\omega(R + r)C)^2 + (1 - (\omega/\omega_0)^2)^2}}. \quad (13)$$

A risonanza un pezzo del denominatore si annulla e si ha $A(\omega = \omega_0) = R/(R + r) = A_{max}$. Osservate che, a causa della presenza della resistenza interna dell'induttore, r , e del fatto che essa comunque "dissipa" potenza, ovvero che si forma un partitore di tensione, questa funzione ha un valore massimo minore di uno, tanto più simile all'unità quanto più r è trascurabile rispetto a R .

Dato che spesso è più pratico esprimere la risposta dell'oscillatore in termini della frequenza f invece che della frequenza angolare ω (è sempre $\omega = 2\pi f$), conviene scrivere esplicitamente anche la funzione $A(f)$:

$$A(f) = \frac{2\pi f RC}{\sqrt{(2\pi f(R + r)C)^2 + (1 - (f/f_0)^2)^2}}, \quad (14)$$

dove $f_0 = 1/(2\pi\sqrt{LC})$ è la frequenza propria dell'oscillatore.

II. CURVA DI RISONANZA E QUALCHE SUA PROPRIETÀ

Scegliamo dei valori verosimili per le grandezze in gioco in modo da poter graficare con Python la $A(f)$, ovvero la curva di risonanza, o spettro, dell'oscillatore; prendiamo $L = 0.5$ H, $C = 0.1$ μ F, $R = 330$ ohm, $r = 40$ ohm. In queste condizioni si ha $f_0 = 712$ Hz. Naturalmente, volendo fare una sorta di simulazione del comportamento del circuito, ci disinteressiamo qui delle incertezze, o tolleranze, sui vari valori impiegati, che quindi vanno intesi come nominali.

Avendo posto dei valori numerici per le grandezze in gioco, possiamo controllare quantitativamente le approssimazioni utilizzate in precedenza. L'approssimazione di sotto-smorzamento è piuttosto ben verificata, dato che $\tau = 2L/R' = 2L/(R + r + r_G) \approx 2$ ms, per cui $1/\tau^2 \approx 2 \times 10^5$ s⁻², mentre $\omega_0^2 = (2\pi f_0)^2 \approx 2 \times 10^7$ (rad/s)². Inoltre possiamo anche notare che sia r_G che r sono minori di R e della impedenza complessiva del circuito a risonanza (che è ovviamente pari a R').

La Fig. 2 mostra il risultato del calcolo di $A(f)$ secondo l'Eq. 14: si osserva una bella campana, centrata su f_0 e di forma *asimmetrica*. Il picco di questa campana, che è ovviamente centrata su $f = f_0$, vale, per la scelta di valori di questo esempio, $R/(R + r) = 0.892$. Il pannello superiore della figura riporta invece lo sfasamento $\Delta\phi$ ottenuto dall'Eq. 12 (calcolando l'arctan di quella espressione): come atteso, lo sfasamento passa per zero cambiando di segno a risonanza, cioè per $f = f_0$, tende a $\pi/2$ per $f \rightarrow 0$ e a $-\pi/2$ per $f \rightarrow \infty$.

Come avrete modo di verificare svolgendo un esercizio (facoltativo), è possibile passare dalla curva di risonanza (nel dominio delle frequenze) all'andamento temporale dell'oscillatore smorzato (nel dominio del tempo) eseguendo una trasformata di Fourier numerica con il metodo FFT.

Un possibile best-fit per i dati sperimentali di $A(f)$, da acquisire in un intervallo di frequenze sufficientemente vasto (tanto più ampio quanto maggiore è il valore di R che avete scelto), può essere condotto usando la funzione di Eq. 14. I parametri da lasciare liberi nel fit dovrebbero essere, in prima battuta, quelli contenenti C , L , r , che non possono essere determinati con sufficiente accuratezza da misure indipendenti. Dunque la funzione di fit potrebbe essere del tipo $g(x) = c_1 x / (\sqrt{(c_2 x)^2 + (1 - (x/c_3)^2)^2})$, con c_{1-3} da determinare. Poiché, però, $c_1 \approx c_2$ (infatti $c_2 = c_1 + 2\pi r C$, con $rC < c_1 = 2\pi RC$), specie nel caso in cui sia stata scelta una resistenza R sufficientemente grande, il numero di parametri liberi del fit potrebbe essere ragionevolmente ridotto a due, in modo da diminuire la covarianza tra di essi e quindi aumentare l'"affidabilità" del best-fit. In alternativa, si potrebbe impostare il valore della frequenza di risonanza f_0 (ovvero il valore del parametro c_3), scegliendolo pari alla frequenza di risonanza misurata in modo diretto, per esempio dalla osservazione dello sfasamento tra V_{in} e V_{out} . Tutte queste possibilità possono

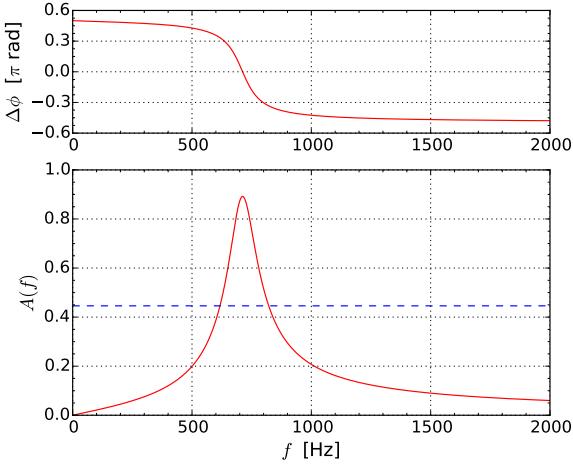


Figura 2. Curva di risonanza calcolata per l'oscillatore descritto nel testo. L'asse verticale riporta il modulo della funzione di trasferimento, $A(f)$, quello orizzontale la frequenza, f . La linea tratteggiata orizzontale rappresenta il valore metà del massimo, $A_{max}/2$. Il pannello superiore mostra lo sfasamento $\Delta\phi$ calcolato per lo stesso oscillatore.

essere esplorate praticamente e confrontate tra loro attraverso il paragone del χ^2 ottenuto (e della covarianza), in modo da stabilire empiricamente quale funzione modello è più adatta per descrivere le osservazioni sperimentali.

Studiamo con un po' di matematica la campana ottenuta. Per caratterizzarne la larghezza Δf usiamo il cosiddetto valore fwhm (*full width at half maximum*, larghezza a metà altezza). La metà altezza $A_{max}/2$ è rappresentata nel grafico dalla linea tratteggiata: possiamo individuare due frequenze, che chiameremo f_{\mp} , che si trovano rispettivamente "a sinistra" e "a destra" di f_0 , per le quali si ha $A(f = f_{\mp}) = A_{max}/2$. Queste frequenze sono dunque quelle a cui il modulo della funzione di trasferimento vale la metà del valore massimo. Dal punto di vista *sperimentale*, tali valori possono essere individuati in maniera piuttosto immediata variando la frequenza del generatore e osservando all'oscilloscopio quando l'ampiezza di V_{out} è la metà di quella di $V_{out,max}$.

Nel fare le misure sul circuito occorre naturalmente tenere conto del fatto che l'ampiezza di V_{in} non rimane costante in tutto l'intervallo di frequenza considerato. Infatti, a risonanza la caduta di potenziale attraverso r_G è non necessariamente trascurabile. Essa può essere determinata considerando che, a risonanza, cioè nelle condizioni che implicano una caduta di potenziale più marcata, l'impedenza complessiva della maglia vale R' . Applicando la regola dei partitori di tensione alla serie costituita da r_G e da "tutto il resto", si ha, per i valori considerati in questo esempio, $|V_{win}| = ((R + r)/R')|V_{\omega G}| \simeq 0.88|V_{\omega G}|$. Dunque il generatore di forme d'onda si comporta, in questo caso, in modo non molto ideale. Infine, sono noti i problemi di termalizzazione che affliggono gli strumenti disponibili in laboratorio, che rendono instabile (general-

mente decrescente) l'ampiezza del segnale in uscita con il passare del tempo, almeno per il primo periodo di funzionamento a freddo. Di conseguenza V_{in} deve essere continuamente monitorata.

In ogni caso, individuate in modo opportuno le frequenze f_- e f_+ , si ha semplicemente $\Delta f_{fwhm} = f_+ - f_-$, che rappresenta un'ottima descrizione (convenzionale) della larghezza della campana. Naturalmente se volessemo esprimere la larghezza fwhm in termini di frequenza angolare avremmo $\Delta\omega_{fwhm} = 2\pi\Delta f_{fwhm}$. Notiamo che sono in uso comune anche altre definizioni per la larghezza di una campana di risonanza. Per esempio spesso si impiega la definizione Δf_{-3dB} , che rappresenta la distanza tra le frequenze in cui V_{out} è $1/\sqrt{2}$ volte il valore massimo $V_{out,max}$, ovvero la distanza tra le frequenze a cui l'attenuazione, o guadagno, $A(f)$ del circuito è -3 dB (ricordate la definizione di attenuazione in dB). Il motivo di questa possibile scelta vi sarà chiaro nel seguito; per il momento limitiamoci a notare che, con un po' di algebra si trova facilmente $\Delta f_{-3dB} \propto \Delta f_{fwhm}$, dove il fattore di proporzionalità, che potete provare a determinare sulla falsariga di quanto verrà esposto tra breve, vale $1/\sqrt{3}$.

Torniamo dunque ad esaminare la larghezza Δf_{fwhm} e facciamo due conticini per vedere dove cadono i valori di f_{\mp} . Si deve risolvere l'equazione algebrica

$$\frac{2\pi f_{\mp} RC}{\sqrt{(2\pi f_{\mp}(R+r)C)^2 + (1 - (f_{\mp}/f_0)^2)^2}} = \frac{1}{2} \frac{R}{R+r}. \quad (15)$$

Facendo il quadrato dei due membri e rimaneggiando si ottiene la seguente equazione:

$$4(2\pi f_{\mp}(R+r)C)^2 = (2\pi f_{\mp}(R+r)C)^2 + (1 - (f_{\mp}/f_0)^2)^2. \quad (16)$$

Poniamo ora $\alpha = (\sqrt{3})2\pi(R+r)C$ e $f_{\mp}^2 = x$. L'espressione precedente può essere riscritta nella forma di un'equazione algebrica di secondo grado per x :

$$x^2 - x(2f_0^2 + \alpha^2 f_0^4) + f_0^4 = 0. \quad (17)$$

Le soluzioni sono:

$$x_{1,2} = \frac{(2f_0^2 + \alpha^2 f_0^4) \pm \sqrt{(2f_0^2 + \alpha^2 f_0^4)^2 - 4f_0^4}}{2} = (18)$$

$$= \frac{(2f_0^2 + \alpha^2 f_0^4)}{2} \pm \alpha f_0^3 \sqrt{1 + \frac{\alpha^2 f_0^2}{4}}. \quad (19)$$

Notiamo che, per la nostra scelta dei componenti, si ha $\alpha \simeq 5 \times 10^{-4}$ s, mentre $f_0 = 712$ Hz. Pertanto la radice quadrata può essere approssimata con l'unità. A questo punto possiamo definire la larghezza (fwhm) della variabile ausiliaria x come $\Delta x_{fwhm} = x_1 - x_2 \approx 2\alpha f_0^3$. Per come questa variabile è stata definita, si ha $\Delta x_{fwhm} \approx 2f_0\Delta f_{fwhm}$. Si trova quindi

$$\Delta f_{fwhm} \approx \alpha f_0^2 = 2\pi\sqrt{3}(R+r)Cf_0^2. \quad (20)$$

Questa è la prima "proprietà" rilevante della curva di risonanza. Per la curva di risonanza graficata in Fig. 2 si

ottiene $\Delta f_{fwhm} = 204$ Hz. Naturalmente questa è anche la larghezza che esce da un'analisi del grafico di Fig. 2, in particolare dalla valutazione della distanza fra le intercette della curva di risonanza con la linea tratteggiata $A_{max}/2$.

Poiché, ricordiamo, $f_0^2 = 1/(2\pi\sqrt{LC})^2$, si ottiene anche

$$\Delta f_{fwhm} \approx \frac{\sqrt{3}}{2\pi} \frac{R+r}{L}; \quad (21)$$

essendo il coefficiente di smorzamento del nostro oscillatore proporzionale al rapporto tra resistenza e induttanza, si verifica quello che già, probabilmente, sapevate dalla meccanica, cioè che *la larghezza della campana di risonanza è proporzionale all'entità dello smorzamento*. In particolare, essa è *inversamente proporzionale* (le dimensioni devono tornare) al tempo di smorzamento $\tau = 2L/R' \simeq 2L/(R+r)$ (supponendo r_G trascurabile rispetto a $R+r$).

C'è poi un'ulteriore "proprietà" rilevante. Si vede facilmente come si abbia $x_1x_2 = f_0^4$, che conduce a

$$f_+ + f_- = f_0^2. \quad (22)$$

Deve sempre verificarsi che il prodotto tra i valori di frequenza ai quali la curva vale un mezzo del valore massimo sia pari al quadrato della frequenza di risonanza. Tutte e due queste "proprietà" sono facilmente verificabili, entro le incertezze sperimentali, nell'esperienza pratica.

Infine osservate che l'asimmetria della campana è tanto minore quanto più il termine $(2f_0^2 + \alpha^2 f_0^4)$ che compare in Eq. 18 è trascurabile rispetto a $\sqrt{(2f_0^2 + \alpha^2 f_0^4)^2 - 4f_0^4}$. Si vede abbastanza facilmente che questo si verifica, e quindi la campana tende a essere simmetrica, quando lo smorzamento è piccolo. Per trovare una relazione matematica che tenga in debito conto delle dimensioni, questo significa che *la campana tende a essere simmetrica per $\alpha f_0 \ll 1$* .

Da ultimo, la Fig. 3 riporta diverse curve di risonanza e i corrispondenti sfasamenti in funzione della frequenza per alcuni diversi valori dei componenti R e C ; si vede come, a parità di L , la frequenza di risonanza sia determinata da C e la larghezza della campana da R .

A. Una visione alternativa

Come succede molto spesso, è possibile giungere qualitativamente alla conclusione che il circuito mostrato in Fig. 1 in modo molto più diretto di quanto consentito dall'applicazione del metodo simbolico. Partiamo dalla considerazione che il segnale V_{out} è rappresentativo della intensità di corrente che circola nella maglia dell'oscillatore. A basse frequenze, tendenti alle condizioni continue ($\omega \simeq 0$), la corrente è "bloccata", o fortemente impedita, dalla presenza del condensatore. Ad alte frequenze, invece, è l'induttore che "reagisce" (per "inerzia") al passaggio di corrente, cercando di impedirlo. Infine, a risonanza,

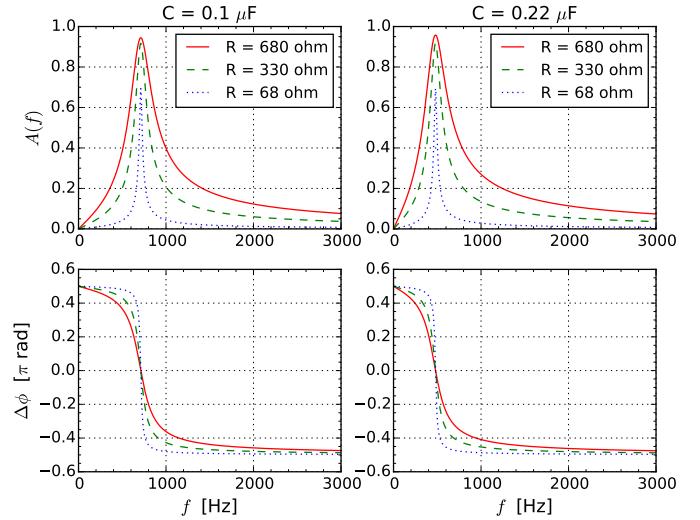


Figura 3. Diverse curve di risonanza e sfasamenti corrispondenti calcolati per diverse scelte di R e C (si suppone $L = 0.5$ H e $r = 40$ ohm per tutte le curve).

come abbiamo già discusso, la serie di condensatore e induttore presenta il minimo di impedenza, che significa la corrente viene massimizzata, ovvero l'ampiezza del segnale V_{out} va al suo picco.

Questo tipo di descrizione può stimolare un approccio, sempre qualitativo, che fa uso dei concetti tipici della descrizione dei filtri. Sappiamo che un filtro, caratterizzato da una data risposta in frequenza, e quindi da una certa funzione di trasferimento, può essere realizzato impiegando resistenze e condensatori. Il comportamento del condensatore (della sua impedenza) con la frequenza è responsabile per la specifica risposta del filtro: usando un solo condensatore e una sola resistenza, a seconda della topologia circuitale, si può ottenere un filtro passa-basso o passa-alto.

È molto semplice dimostrare che anche unendo un induttore e una resistenza in una maglia si può ottenere un filtro. Poiché il comportamento dell'induttore con la frequenza è "opposto" a quello del condensatore (il condensatore tende a comportarsi da circuito aperto o da cortocircuito a frequenze rispettivamente basse o alte, l'induttore tende a comportarsi in modo opposto), è altrettanto facile rendersi conto che la topologia di un filtro passa-alto RL è la stessa di un passa-basso RC con l'induttore al posto del condensatore, quella di un filtro passa-basso RL è la stessa di un passa-alto RC con l'induttore al posto del condensatore.

Nel circuito di Fig. 1 il condensatore è in serie all'induttore e inoltre è presente una resistenza verso la linea di terra, o massa. Con un po' di forzatura concettuale, quel circuito può essere pensato come la sequenza di un filtro passa-alto (in cui l'elemento che reagisce alla frequenza è il condensatore) e di un filtro passa-basso (in cui invece è protagonista l'induttore). La serie dei due filtri

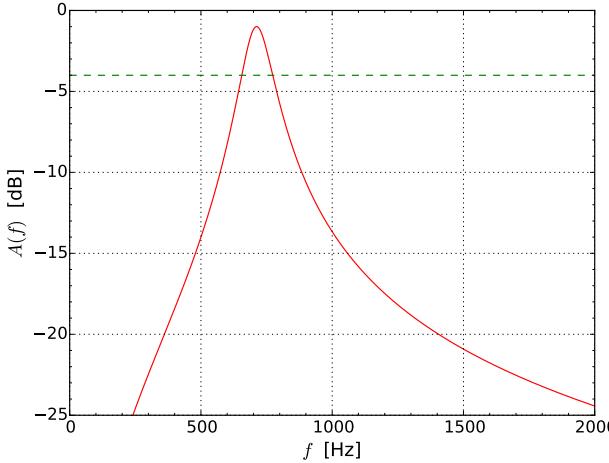


Figura 4. Analogo del pannello inferiore di Fig. 2, ma con l’attenuazione, o guadagno, $A(f)$ misurata in dB, come si fa nel diagramma di Bode dei filtri. La linea tratteggiata orizzontale rappresenta il valore che si trova a -3 dB rispetto al valore massimo. Si può facilmente dimostrare che la larghezza della campana così determinata, cioè l’intervallo in frequenza tra le due intercette della linea con la campana, è $\Delta f_{-3dB} = \Delta f_{fwhm}/\sqrt{3}$.

dà luogo a un filtro *passa-banda*, che lascia passare, cioè non attenua (o attenua di poco) le frequenze comprese all’interno di un certo range, che in pratica corrisponde alla larghezza della campana della curva di risonanza.

In linea di principio, un filtro passa-banda potrebbe anche essere costruito usando solo resistenze e condensatori (o, se si vuole, resistenze e induttori). Tuttavia la variante *RLC* è sicuramente preferibile: (i) essa minimizza i problemi connessi con il montaggio in serie di due distinti circuiti, e la conseguente necessità di aggiustare l’impedenza di uscita del primo sotto-circuito con quella di ingresso del sotto-circuito seguente; (ii) a patto di usare valori di resistenza sufficientemente bassi, essa permette di ottenere bande relativamente strette; (iii) il comportamento con la frequenza di un oscillatore forzato *RLC* permette di ottenere curve di attenuazione, o guadagno, che sono generalmente più “ripide” di quelle realizzate per filtri “a un polo”.

Per curiosità, la Fig. 4 mostra la curva di risonanza (parametri come in Fig. 2) con il modulo della funzione di trasferimento, $A(f)$, misurato in unità di dB, come si fa quando si preparano i grafici di Bode dei filtri. Si vede chiaramente come, almeno in alcuni tratti, la pendenza della curva (in valore assoluto) sia ben maggiore dei 3 dB/ottava (o 20 dB/decade) tipici dei filtri “a un polo”. Inoltre, come già affermato, diminuendo la resistenza della maglia potrebbero facilmente essere ottenute pendenze ancora più marcate.

III. ENERGIA, POTENZA E FATTORE DI QUALITÀ (Q-FACTOR)

L’aspetto più trasversale del fenomeno della risonanza nelle sue molteplici applicazioni riguarda il comportamento del sistema, che per noi è un oscillatore *RLC*, nei confronti della potenza. Infatti si afferma spesso che a risonanza si ottiene il massimo trasferimento di potenza dalla forzante, che per noi è il generatore di forme d’onda, al sistema. Scopo di questa sezione è cercare una descrizione quantitativa di questo fenomeno soffermandoci su diversi aspetti che riguardano energia, potenza media, e fattore di qualità dell’oscillatore.

Come sappiamo dallo studio dell’oscillatore smorzato *RLC*, nella descrizione di un circuito in cui induttore e condensatore si trovano in serie fra loro è possibile individuare due tipologie, ovvero due espressioni, per l’energia. Essa, infatti, può avere sede sia nel condensatore, e quindi avere un carattere elettrostatico, che nell’induttore, e quindi avere un carattere magnetostatico: sappiamo anche che tutto ciò funziona in analogia con l’oscillatore meccanico, dove l’energia può essere espressa come elastica e cinetica. Anche nell’oscillatore *RLC* l’energia passa continuamente e periodicamente da elettrostatica a magnetostatica, così come nell’oscillatore meccanico essa passa continuamente e periodicamente da elastica a cinetica (per chi conosce il concetto, c’è un vettore di Poynting che punta da induttore a condensatore, e viceversa, in modo oscillante nel tempo).

Nel funzionamento dell’oscillatore si ripetono periodicamente degli istanti di tempo in cui la carica sul condensatore è nulla e la corrente che circola nella maglia è massima (al valore I_{max}), cioè in cui l’energia è espressa solo dal termine magnetostatico. L’espressione di questa energia, che quindi rappresenta *tutta* l’energia immagazzinata nell’oscillatore, è

$$E_{\text{stored}} = \frac{L}{2} I_{max}^2. \quad (23)$$

Ora è facile rendersi conto che $I_{max}^2 \propto |V_{out}|^2$: infatti fare il quadrato dell’ampiezza della corrente massima equivale a considerare il modulo quadro del fasore I_ω corrispondente, che è sicuramente proporzionale, attraverso l’inverso del modulo dell’impedenza di uscita del circuito, $|Z_{out}|$, a V_{out} . Ora, supponendo di trovarci in condizioni in cui $|V_{in}|$ rimane costante mentre viene variata la frequenza e costruita la curva, o spettro, di risonanza, risulta evidente che

$$E_{\text{stored}}(f) \propto A^2(f), \quad (24)$$

dove rinunciamo a esplicitare il coefficiente di proporzionalità, che qui non ci interessa.

Dunque, poiché l’andamento dell’energia immagazzinata nell’oscillatore dipende da $A^2(f)$, spesso è utile avere una rappresentazione grafica proprio di questa funzione, a cui possiamo dare il nome di *curva di risonanza*, o *spettro di energia*. In Fig. 5, ad esempio, il calcolo di

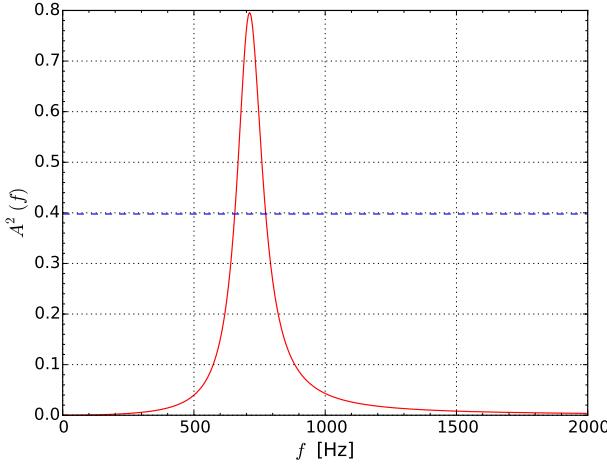


Figura 5. Calcolo della curva di risonanza, o spettro, di energia $A^2(f)$ per lo stesso oscillatore armonico RLC considerato in Fig. 2. La linea tratteggiata orizzontale rappresenta il valore $A^2_{max}/2$. Si può facilmente dimostrare che la larghezza a metà altezza della campana così determinata, cioè l'intervallo in frequenza tra le due intercette della linea con la campana, è $\Delta f_{fwhm,A^2} = \Delta f_{-3dB}$.

$A^2(f)$ è eseguito per la stessa scelta di valori di Fig. 2. Si vede come, per effetto del quadrato, la campana sia più stretta e più simmetrica rispetto a quella di Fig. 2. Inoltre, come è facile dimostrare, la sua larghezza a metà altezza è $\Delta f_{fwhm,A^2} = \Delta f_{-3dB}$, che dà ragione del perché sia comune definire e impiegare anche la larghezza Δf_{-3dB} per caratterizzare una curva di risonanza.

Concediamoci un'altra piccola divagazione matematica sulla $A^2(f)$. Troverete in futuro, se già non è successo, che alcuni spettri, o curve di risonanza, hanno una forma ben descritta da una funzione Lorentziana, cioè del tipo

$$h(x) = \frac{\kappa_1}{\kappa_2 + (x - x_0)^2}, \quad (25)$$

con κ_1 e κ_2 costanti e x variabile indipendente generica. Questa funzione rappresenta una campana *simmetrica*, centrata in $x = x_0$, con altezza di picco $h_{max} = \kappa_1/\kappa_2$ e larghezza a metà altezza $\Delta f_{fwhm} = 2\sqrt{\kappa_2}$. Confrontata con una Gaussiana di simile larghezza, la Lorentziana ha l'importante caratteristica di avere delle code "più alte".

Si può dimostrare che, nel caso di oscillatore RLC molto poco smorzato, la $A^2(f)$ tende a questa forma funzionale. Scriviamo infatti esplicitamente la $A^2(f)$, cioè svolgiamo il quadrato dell'Eq. 14:

$$A^2(f) = \frac{(2\pi f RC)^2}{(2\pi f(R+r)C)^2 + (1 - (f/f_0)^2)^2} \simeq \quad (26)$$

$$\simeq \frac{(2\pi f RC)^2}{(2\pi f RC)^2 + (1 - (f/f_0)^2)^2}, \quad (27)$$

dove l'ultimo passaggio vale per $r \ll R$, che supponiamo valida per semplicità matematica. Con alcuni, ulteriori e

semplicissimi, passaggi, si ha

$$A^2(f) \simeq \frac{1}{1 + \frac{(f_0^2 - f^2)^2}{(2\pi f RC)^2 f_0^4}} = \quad (28)$$

$$= \frac{1}{1 + \frac{((f_0 - f)(f_0 + f))^2}{(2\pi f RC)^2 f_0^4}}. \quad (29)$$

Se l'oscillatore è molto poco smorzato, allora la larghezza della campana $A^2(f)$ sarà molto stretta, cioè la funzione sarà sensibilmente diversa da zero solo in un piccolo intorno di $f = f_0$. A parte che nel termine $(f - f_0)$, in cui conta (al primo ordine) la differenza tra f e f_0 , altrove potremo porre $f \simeq f_0$, ottenendo

$$A^2(f) \simeq \frac{1}{1 + 4 \frac{(f_0 - f)^2}{f_0^4} \frac{1}{(2\pi f RC)^2}}. \quad (30)$$

Questa è proprio l'espressione di una Lorentziana centrata in $f = f_0$, cioè sulla risonanza, e di larghezza a metà altezza $\Delta f_{fwhm,A^2} = 2\pi f_0 RC = R/(2\pi L)$, dove abbiamo usato $f_0^2 = 1/(4\pi^2 LC)$. Ricordando che per il nostro oscillatore armonico il tempo di smorzamento è $\tau = 2L/R$, la larghezza a metà altezza si può scrivere come $\Delta f_{fwhm,A^2} = 1/(\pi\tau)$, ovvero, tornando a ragionare in termini di frequenza angolare, $\Delta\omega_{fwhm,A^2} = 2/\tau$.

Al di là dei dettagli matematici, che non contano molto, il messaggio di questa divagazione può essere riassunto così: per un oscillatore molto poco smorzato, cioè per il quale $1/\tau \ll \omega_0$, la curva di risonanza di energia tende ad assumere la forma Lorentziana, con una semilarghezza a metà altezza (misurata in unità di frequenza angolare) pari a $1/\tau$. Per esempio, questo è il caso del cosiddetto "modello di Lorentz" dell'interazione radiazione/materia in approccio classico, ma anche l'andamento di una popolazione di sistemi (per esempio atomi) eccitati, la cui eccitazione viene smorzata attraverso collisioni (questo esempio può anche essere messo in relazione con la "distribuzione di Cauchy" da voi già studiata in precedenza).

Ragioniamo ora in termini di potenza. Sappiamo che, in un circuito sottoposto a una d.d.p. alternata (sinusoidale) $\Delta V(t)$ e attraversato da una corrente alternata (sinusoidale) di intensità $I(t)$, la potenza media può essere espressa come $\langle P \rangle = (\Delta V_{max} I_{max}/2) \cos(\Delta\phi)$, dove ΔV_{max} e I_{max} rappresentano le ampiezze (qui considerate reali) dei segnali $\Delta V(t)$ e $I(t)$, ovvero i moduli dei corrispondenti fasori, e $\Delta\phi$ lo sfasamento dell'uno rispetto all'altro. Come certamente ricordate, sia il fattore $1/2$ che il fattore di potenza $\cos(\Delta\phi)$ vengono dall'operazione di media temporale.

Se consideriamo la potenza erogata dal generatore di forme d'onda che alimenta il circuito RLC , possiamo individuare $\Delta V(t)$ con il segnale $V_{in}(t)$ [o, se preferite, con $V_G(t)$] e $I(t)$ con la corrente che scorre nella maglia dell'oscillatore. Sulla falsariga di quanto esposto per il calcolo dell'energia immagazzinata nell'oscillatore, possiamo facilmente dedurre che $\langle P(f) \rangle \propto A^2(f) \cos(\Delta\phi)$. Visto

l'andamento di $\Delta\phi$ con la frequenza f , è immediato affermare che questa funzione è ancora una campana centrata in $f = f_0$. Se interpretiamo $\langle P(f) \rangle$ come potenza *trasferita* dalla forzante al sistema, è evidente che essa ha un picco proprio a risonanza, dove lo sfasamento è nullo.

A risonanza, nell'oscillatore l'energia rimbalza continuamente e periodicamente tra condensatore (energia elettrostatica) e induttore (energia magnetostatica). La presenza delle resistenze, che questa volta consideriamo in blocco come $R' = R + r + r_G$, implica dissipazione di potenza per effetto Joule. La potenza *media* dissipata per effetto Joule è

$$\langle P_{Joule} \rangle = \frac{R'I_{max}^2}{2}. \quad (31)$$

A risonanza, questa potenza, che l'oscillatore perde (cioè “dissipa” mediamente nel tempo) viene mediamente ripianata dalla potenza erogata dal generatore, trasferita continuamente al sistema.

Come già fatto nello studio dell'oscillatore smorzato *RLC*, possiamo a questo punto introdurre una grandezza che misura l'*energia persa per ciclo*:

$$E_{lostpercycle} = \langle P_{Joule} \rangle T_0 = \frac{R'I_{max}^2}{2f_0}, \quad (32)$$

dove $T_0 = 1/f_0$ rappresenta il periodo dell'oscillazione a risonanza. Combinando energia immagazzinata e energia persa nel ciclo, possiamo scrivere il *fattore di qualità* come

$$Qf = 2\pi \frac{E_{stored}}{E_{lostpercycle}}, \quad (33)$$

da cui

$$Qf = 2\pi \frac{LI_{max}^2/2}{R'I_{max}^2/2} f_0 = 2\pi \frac{L}{R'} f_0. \quad (34)$$

È facile notare allora che è dimostrata la relazione $Qf \propto f_0/\Delta f_{fwhm}$ (qui supponiamo r_G trascurabile); il fattore di proporzionalità, che vale $\sqrt{3}$, cioè $Qf = \sqrt{3}f_0/\Delta f_{fwhm}$, è legato alla definizione di larghezza a metà altezza (vedi Eq. 21). In particolare, esso scomparirebbe se usassimo la larghezza Δf_{-3dB} che abbiamo definito prima.

Di conseguenza, l'analisi della curva di risonanza, in particolare di quella per l'energia, conduce in maniera pressoché immediata alla determinazione di Qf per un oscillatore armonico.

Per concludere questa sezione, facciamo due conticini: per l'oscillatore costruito con i componenti citati nel nostro esempio si ha $Qf \approx 6$, un valore molto inferiore a quello che si ottiene sopprimendo la resistenza R , come nell'esperienza con l'oscillatore smorzato.

IV. CIRCUITO “ANTIRISONANTE”

Come potete facilmente capire, i circuiti con resistori, induttori, condensatori sono un ottimo banco di prova per

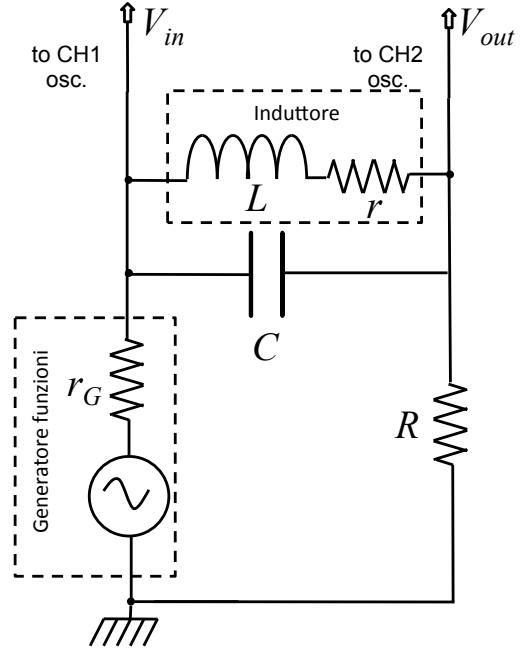


Figura 6. Circuito dell'oscillatore “antirisonante” (parallelo) considerato nel testo.

verificare, studiare, analizzare le condizioni di risonanza. A differenza degli analoghi meccanici, qui si può godere di un'ampia libertà nel definire le condizioni di operazione del circuito (sempre a patto che le approssimazioni considerate si mantengano valide).

Un esempio della versatilità ed efficacia dei circuiti di questo tipo è rappresentato dallo schema di Fig. 6, che rappresenta quello che talvolta si chiama *oscillatore antirisonante*. La denominazione intende mettere in luce che, in tale circuito, a risonanza l'ampiezza di V_{out} diminuisce, cioè ha un picco orientato “verso il basso” laddove prima questo era orientato “verso l'alto”.

Esaminiamo brevemente anche questo circuito. La differenza fondamentale rispetto al precedente è che stavolta induttore e condensatore sono in parallelo fra loro. L'impedenza di questo parallelo è

$$Z_{par} = \frac{r + j\omega L}{j\omega rC + (1 - \omega^2 LC)}. \quad (35)$$

La soluzione completa, cioè senza approssimazioni, dell'equazione del circuito è in questo caso parecchio complicata: essa è trattata numericamente in Appendice. Per proseguire con un po' di matematica semplice, occorre imporre che sia trascurabile la resistenza interna dell'induttore rispetto a ωL , cioè $r \ll \omega L$. Tenendo conto dei valori in gioco, questa approssimazione è attesa valere abbastanza bene solo per frequenze sufficientemente alte.

Nell'approssimazione fatta, e assumendo per semplicità anche r_G trascurabile, per cui $V_{\omega G} = V_{win}$, si ottiene abbastanza rapidamente la seguente funzione di

trasferimento:

$$T(\omega) = \frac{V_{\omega out}}{V_{\omega in}} = \frac{j\omega rC + (1 - (\omega/\omega_0)^2)}{j\omega(L/R + rC) + (1 - (\omega/\omega_0)^2)}, \quad (36)$$

con, come al solito, $\omega_0 = 1/\sqrt{LC}$. In questo circuito la risonanza implica che l'impedenza del parallelo tenda a un massimo; di conseguenza a risonanza diminuisce la corrente che passa nel circuito e quindi V_{out} ha un minimo.

Infatti il modulo della funzione di trasferimento, scritto anche stavolta in funzione di $f = \omega/(2\pi)$, è

$$A(f) = |T(f)| = \sqrt{\frac{(2\pi f r C)^2 + (1 - (f/f_0)^2)^2}{(2\pi f(rC + L/R))^2 + (1 - (f/f_0)^2)^2}}. \quad (37)$$

Questa funzione è decisamente meno comprensibile di quella espressa in Eq. 14, però ci si rende conto abbastanza facilmente che essa presenta un *minimo* per $f = f_0$. La Fig. 7 mostra un grafico della funzione $A(f)$ per valori dei componenti analoghi a quelli usati in precedenza. Si tralasciano le verifiche matematiche, ma si può osservare come anche in questo caso si ottenga che a risonanza lo sfasamento passi per zero e cambi di segno. Inoltre anche qui la larghezza della campana dipende dal coefficiente di smorzamento (ma, attenzione, la dipendenza con il coefficiente di smorzamento è opposta rispetto a prima, cioè diminuisce con R' , l'espressione di Δf_{fwhm} è diversa e la larghezza della campana è in genere ben maggiore, come si vede anche dal grafico), e valgono alcune "proprietà" simili a quelle trovate prima per l'oscillatore risonante. In particolare si ha ancora $f_- f_+ = f_0^2$. Notate che l'andamento "spigoloso" e il quasi annullamento della funzione per $f \approx f_0$ possono essere, almeno in parte, conseguenza delle approssimazioni fatte.

APPENDICE: CALCOLO DI $A(f)$ PER L'OSCILLATORE ANTI-RISONANTE CON IL PACCHETTO CMATH

Come sappiamo, Python dispone di un pacchetto, denominato **cmath**, che consente di manipolare numeri complessi. Adoperare i numeri complessi può tornare utile quando la funzione $T(f)$ diventa complicata, come nel caso del circuito antirisonante. Infatti, almeno in linea di principio, non c'è bisogno di agire con razionalizzazioni, semplificazioni e altri artifici matematici per ottenere (sulla carta) una funzione $A(f)$ che possa essere direttamente calcolata con Python.

Scriviamo l'impedenza totale del circuito come

$$Z_{tot} = r_G + R + Z_{par}, \quad (38)$$

con

$$Z_{par} = \left(\frac{1}{r + j\omega L} + j\omega C \right)^{-1}, \quad (39)$$

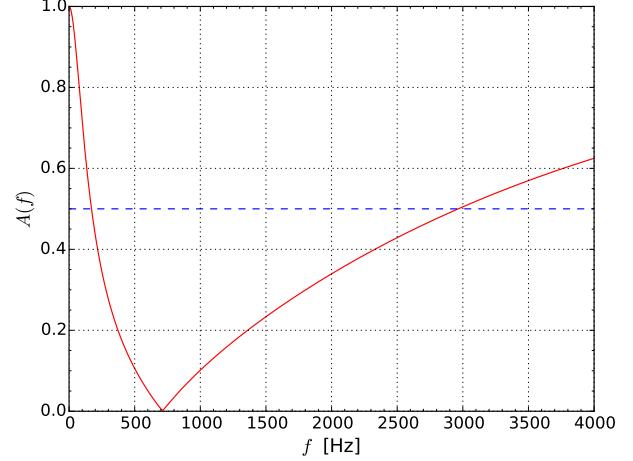


Figura 7. Curva di "antirisonanza" calcolata per l'oscillatore (parallelo) descritto nel testo. L'asse verticale riporta il modulo della funzione di trasferimento, $A(f)$, quello orizzontale la frequenza, f . La linea tratteggiata orizzontale rappresenta il valore metà del massimo, $A_{max}/2$.

impedenza del parallelo induttore reale e condensatore, scritta in forma complessa. Questa espressione è ovviamente analoga a quella di Eq. 35, ma, coerentemente con l'approccio che stiamo seguendo, in essa non sono stati inseriti passaggi matematici di alcun genere.

Il segnale di uscita, ovvero il fasore $V_{\omega out}$, è preso ai capi del resistore R , mentre il segnale di ingresso, $V_{\omega in}$, è preso ai capi della serie costituita dal resistore R , con impedenza R , e dal parallelo tra induttore (reale) e condensatore, con impedenza Z_{par} . Possiamo quindi porre

$$Z_{out} = R \quad (40)$$

$$Z_{in} = R + Z_{par}. \quad (41)$$

La funzione di trasferimento può essere convenientemente espressa in funzione delle impedenze appena definite. Infatti, dato che $V_{\omega out} = Z_{out}I_\omega$ e $V_{\omega in} = Z_{in}I_\omega$, si può scrivere

$$T(\omega) = \frac{Z_{out}}{Z_{in}}. \quad (42)$$

Usando Python, è facile costruire arrays di Z_{out} e Z_{in} in funzione della frequenza f ed è immediato calcolarne i moduli. Dividendo questi moduli si ottiene sotto forma di array la funzione di trasferimento $A(f)$, mentre calcolando il rapporto tra parte immaginaria e reale si ha lo sfasamento (che quindi è anch'esso semplice da calcolare).

Il risultato del calcolo, sia per $A(f)$ che per $\Delta\phi$, è mostrato in Fig. 8: i valori dei componenti del circuito sono quelli già impiegati in precedenza, cioè $R = 330$ ohm, $r = 40$ ohm, $r_G = 50$ ohm, $L = 0.5$ H, $C = 0.1$ μ F. Si vede che effettivamente lo sfasamento passa per lo zero,

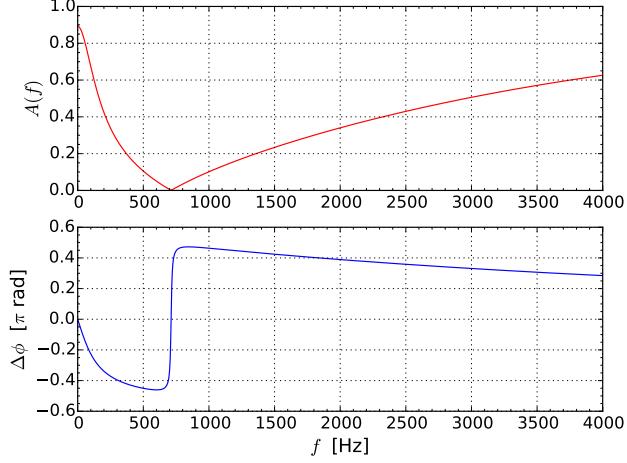


Figura 8. Attenuazione, o guadagno, $A(f)$ e sfasamento $\Delta\phi$ in funzione della frequenza per il circuito antirisonante di Fig. 6, calcolate per i valori dei componenti citati nel testo. Il calcolo è eseguito usando direttamente le espressioni complesse delle impedenze in gioco, grazie all'impiego del pacchetto `cmath` di Python.

cambiando di segno, a risonanza. Per quanto riguarda l'attenuazione, o guadagno, l'andamento è simile a quello approssimato di Fig. 7, con un po' meno spigolosità a basse frequenze, a testimonianza che le approssimazioni introdotte per ricavare l'Eq. 37 sono ragionevoli.

Auto e mutua induzione

francesco.fuso@unipi.it

(Dated: version 7 - FF, 2 aprile 2022)

Questa breve nota affronta l'argomento della mutua induzione, con qualche piccolo riferimento all'attività sperimentale. Essa si basa sul materiale realizzato nei bei periodi di esercitazione in presenza e senza restrizioni, quando era possibile svolgere esperimenti finalizzati alla misura diretta delle grandezze rilevanti. Pertanto alcuni accenni potrebbero risultare poco coerenti rispetto a quanto fatto quest'anno.

I. MUTUA INDUZIONE

Campi di induzione magnetica variabili nel tempo presenti in una qualche regione di spazio possono “accoppiarsi” a circuiti elettrici dando luogo a d.d.p. variabili nel tempo su questi circuiti. È infatti sufficiente che il circuito elettrico definisca una superficie attraverso la quale la variazione del flusso del campo di induzione magnetica sia non nullo; in queste condizioni l'equazione di Faraday mostra che viene indotta un d.d.p. variabile nel tempo, a cui, in circuiti elettrici ordinari, corrisponde una corrente variabile nel tempo. L'induzione magnetica è sicuramente ben nota nei suoi effetti. Ad esempio, anche l'ubiquo *pick-up noise*, spesso e volentieri presente nelle misure svolte in laboratorio didattico (dove non si usano, o quasi, cavi coassiali), può essere visto come una conseguenza dell'induzione magnetica.

Gli esperimenti di nostro interesse sono progettati per verificare l'accoppiamento magnetico tra circuiti con geometria ben definita. Nella sostanza, infatti, vengono usate delle bobine, o induttori, e si studia l'accoppiamento tra diversi induttori in diverse configurazioni, supponendo trascurabili tutti gli altri effetti spuri.

Usiamo degli indici ij per indicare una coppia di questi induttori. Supponiamo che l'induttore i -esimo sia percorso da una corrente variabile nel tempo $I_i(t)$: essa produce quindi un campo di induzione magnetica $\vec{B}_i(t)$ variabile nel tempo che insiste in una certa regione di spazio, tipicamente non delimitata. In questa regione di spazio si trova l'induttore j -esimo ed esiste una superficie delimitata dall'avvolgimento dell'induttore j -esimo su cui il flusso del campo $\vec{B}_i(t)$ è non nullo. Indichiamo questo flusso come $\Phi_j(\vec{B}_i(t))$.

Definiamo *coefficiente di mutua induzione*, o *mutua induttanza*, della coppia di induttori ij la grandezza

$$M_{ji} = \frac{\Phi_j(\vec{B}_i(t))}{I_i(t)}. \quad (1)$$

È evidente che il coefficiente di autoinduzione L , quello che abbiamo già incontrato e utilizzato in varie occasioni, è un elemento diagonale della matrice M_{ji} , ovvero

$$L_i = M_{ii}. \quad (2)$$

Inoltre, come si può dimostrare in maniera rigorosa e come ci accontentiamo di intuire sulla base di argomenti di

reciprocità (gli indici i e j possono scambiarsi tra loro senza modificare la configurazione geometrica del problema), ripromettendoci comunque di tornare sull'argomento in altra sede, si ha

$$M_{ji} = M_{ij}, \quad (3)$$

ovvero la matrice M_{ji} è simmetrica.

La forza elettromotrice indotta dalla variazione di flusso campo $\vec{B}_i(t)$ sull'induttore j -esimo si scrive, usando la legge di Faraday

$$\varepsilon_j = -\frac{d\Phi_j(\vec{B}_i(t))}{dt}, \quad (4)$$

che, impiegando l'Eq. 1, diventa

$$\varepsilon_j = -M_{ji} \frac{dI_i(t)}{dt}. \quad (5)$$

La forza elettromotrice si traduce in una d.d.p. indotta; la “relazione costitutiva” della mutua induzione, che lega la d.d.p. ΔV_j sull'induttore j -esimo alla variazione dell'intensità di corrente $I_i(t)$ sull'induttore i -esimo, diventa quindi, in analogia con il trattamento riservato all'autoinduzione,

$$\Delta V_j = M_{ji} \frac{dI_i(t)}{dt}. \quad (6)$$

Noteate che nel contesto di cui ci stiamo occupando, come sarà chiaro anche in alcune delle applicazioni considerate nel seguito, la scelta dei segni è necessariamente legata a delle convenzioni. I segni delle correnti che circolano nelle due bobine sono concordi solo quando gli avvolgimenti sono realizzati in verso concorde rispetto a uno stesso riferimento.

La ΔV_j implica a sua volta una corrente indotta $I_j(t)$ nel circuito dell'induttore j -esimo. Gli effetti possono essere trattati sia nel dominio del tempo che in quello delle frequenze. Per praticità, in questa nota ci limitiamo allo studio nel dominio delle frequenze, supponendo segnali *sinusoidali* come quelli usati in laboratorio e quindi usando il metodo simbolico: siete invitati a pensare da soli alla trattazione nel dominio del tempo. Inoltre, tenendo conto che in questa nota ci restringiamo a considerare solo due induttori (gli indici i e j possono essere solo 1 o 2), per semplificare la notazione poniamo $M_{11} = L_1$, $M_{22} = L_2$, $M_{21} = M_{12} = M$.

A. Energia e mutua induzione

La presenza della mutua induzione comporta anche un contributo specifico nell'energia di configurazione ("magnetica") del sistema dei due induttori. Supponiamo che essi siano percorsi da correnti di intensità I'_1 e I'_2 e chiamiamo ΔV_1 e ΔV_2 le d.d.p. ai loro capi. La potenza associata al passaggio di corrente può essere espressa come $P_{1,2} = \Delta V_{1,2} I'_{1,2}$. Indichiamo i valori "a regime" delle intensità di corrente come I_1 e I_2 e individuiamo una semplice strategia che permetta alle correnti di passare da zero, situazione in cui l'energia magnetica è nulla, a questi valori di regime. Supponiamo per esempio una prima fase in cui $I'_2 = 0$ (costante, per esempio mantenendo aperto il circuito a cui è collegata la bobina 2) e $I'_1 = 0 \rightarrow I_1$: in questa fase la potenza sulla seconda bobina è nulla e l'unico contributo all'energia è quello della bobina 1, che si ottiene integrando nel tempo la P_1 portando a $U_{M,1} = L_1 I_1^2 / 2$. Nella fase successiva supponiamo di mantenere costante al valore I_1 l'intensità della corrente che scorre nella prima bobina e di avere $I'_2 = 0 \rightarrow I_2$. In questa fase viene attivato il contributo all'energia magnetica dovuto alla corrente I_2 , $U_{M,2} = L_2 I_2^2 / 2$. Inoltre la variazione della d.d.p. ai capi della bobina 1 dovuta alla mutua induzione, $\Delta V_M = M dI_2 / dt$, implica un termine addizionale di potenza $P_1 = \Delta V_2 I_1 = M I_1 dI_2 / dt$. Questa potenza, integrata nel tempo, fornisce un ulteriore contributo all'energia dovuto alla mutua induzione: $U_{M,1,2} = M I_1 I_2$. L'energia magnetica complessiva è quindi

$$U_M = \frac{L_1 I_1^2}{2} + \frac{L_2 I_2^2}{2} + M I_1 I_2 . \quad (7)$$

In generale, per un sistema in cui ci sono tante bobine e la mutua induzione è descritta dalla matrice M_{ij} , è $U_M = \sum_{ij} M_{ij} I_i I_j / 2$.

B. Equazioni del primario e del secondario

Usando una terminologia tipica del mondo dei trasformatori, e quindi anticipando un argomento che tratteremo in seguito, indichiamo i circuiti che comprendono le bobine 1 e 2 come rispettivamente *primario* e *secondario*. Supponiamo poi che in questi circuiti, oltre agli elementi reattivi di auto e mutua induzione, ci siano solo elementi resistivi, indicati con R_1 e R_2 , eventualmente comprendenti anche le resistenze interne degli induttori, e indichiamo i fasori delle d.d.p. sui due circuiti come $V_{\omega 1}$ e $V_{\omega 2}$.

Le equazioni dei circuiti primario e secondario nel dominio delle frequenze (segnali sinusoidali) si scrivono

$$V_{\omega 1} = (R_1 + j\omega L_1) I_{\omega 1} + j\omega M I_{\omega 2} \quad (8)$$

$$V_{\omega 2} = (R_2 + j\omega L_2) I_{\omega 2} + j\omega M I_{\omega 1} , \quad (9)$$

dove abbiamo tenuto in debito conto l'effetto della mutua induzione. Infatti, per fare un esempio, a causa della

mutua induzione il fasore di corrente $I_{\omega 1}$ responsabile per la creazione del campo $\vec{B}_1(t)$ induce una d.d.p. variabile nel tempo nel circuito 2, che si scrive $j\omega M I_{\omega 1}$. Dunque l'accoppiamento magnetico tra primario e secondario si riflette, in generale, nell'accoppiamento tra le equazioni dei due circuiti.

Il coefficiente di mutua induzione M è determinato fondamentalmente dalla geometria del sistema (forma, dimensioni, numero spire degli avvolgimenti, orientazione reciproca degli assi degli avvolgimenti, loro distanza reciproca, eventuale presenza di materiali ferromagneticici, etc.). In parte, questi parametri sono anche quelli che determinano i coefficienti di autoinduzione, L_1 e L_2 . Per motivi che risulteranno più chiari nel seguito della nota, può convenire esprimere M in funzione di L_1 e L_2 tramite la

$$M = k \sqrt{L_1 L_2} , \quad (10)$$

dove k è un numero (dimensioni e unità di misura sono le stesse per L e M), detto *coefficiente di accoppiamento magnetico*, compreso tra 0 e 1 che misura l'"efficacia" con cui le linee di campo di induzione magnetica prodotte da una bobina sono concatenate con la sezione degli avvolgimenti dell'altra. Infatti è evidentemente $k \simeq 0$ nel caso in cui $M \simeq 0$, essendo generalmente $L_{1,2} \neq 0$.

Verifichiamo ora che deve essere sempre $k \leq 1$ e che $k = 1$ implica accoppiamento *completo*, cioè che *tutte* le linee di campo di induzione magnetica che passano per la sezione dell'induttore 1 sono anche concatenate con la sezione dell'induttore 2. A questo scopo dividiamo il secondo membro dell'Eq. 7 per I_2^2 e poniamo $x = I_1 / I_2$; otteniamo quindi la seguente equazione algebrica di secondo grado in x :

$$g(x) = \frac{L_1}{2} x^2 + Mx + \frac{L_2}{2} , \quad (11)$$

che non può essere mai negativa, essendo impossibile che $U_M < 0$. L'Eq. 11 ha soluzioni $x_{1,2} = (-M \pm \sqrt{\Delta}) / L_2$, con $\Delta = (M^2 - L_1 L_2)$. È facile verificare che $g(x) \geq 0$ implica $\Delta \leq 0$, ovvero, tenendo conto della definizione di Eq. 10, $k \leq 1$. Inoltre la condizione $g(x) = 0$ significa che esiste un dato valore del rapporto x ($x = -M / L_2$) tra le intensità di corrente nel primario e secondario tale che l'energia magnetica totale si annulla. Poiché la *densità* di energia magnetica è proporzionale al quadrato dei campi, $g(x) = 0$ vuol dire fisicamente che, per un dato valore di x dipendente da auto e mutua induzione del sistema considerato, il campo di induzione magnetica è *nullo in tutto lo spazio*. Ciò è possibile solo nel caso di accoppiamento completo, quando facendo circolare una certa intensità di corrente in un induttore si ottiene l'annullamento del campo di induzione magnetica all'interno dell'altro induttore.

II. NUCLEI FERROMAGNETICI E CANALIZZAZIONE DELLE LINEE DI CAMPO

In questa sezione consideriamo gli effetti dovuti alla presenza di un materiale *ferromagnetico*, cioè con permeabilità magnetica relativa $\mu_r >> 1$, nella regione di spazio in cui è presente un campo di induzione magnetica prodotto da una bobina, per esempio una configurazione in cui il materiale si trova nel nucleo dell'induttore. Gli argomenti qui considerati non sono rilevanti negli esperimenti svolti in laboratorio quest'anno, ma comunque è opportuno nella trattazione supponiamo valida la relazione tra campi magnetici

$$\vec{B} = \mu_0 \mu_r \vec{H}, \quad (12)$$

che implica *omogeneità*, *isotropia*, *linearità*; torneremo in futuro a discutere sulla validità e sugli aspetti fisici di queste assunzioni. Qui ci limitiamo a osservare che, a parità delle altre condizioni che determinano il campo magnetico \vec{H} in una certa configurazione sperimentale, il campo di induzione magnetica dentro il materiale risulta amplificato in intensità.

Abbiamo già avuto modo di commentare come, a causa del loro carattere generalmente conduttore (esistono tuttavia eccezioni notevoli dal punto di vista tecnologico, ad esempio le *ferriti*), i materiali ferromagnetici possano essere sede di *correnti parassite* indotte da campi di induzione magnetica variabili nel tempo, in modo qualitativamente simile a qualsiasi materiale conduttore. Oltre ad accrescere i fenomeni dissipativi, le correnti parassite possono agire sull'intensità del campo di induzione magnetica attraverso meccanismi di "schermatura" a cui può conseguire una diminuzione del coefficiente di autoinduzione e una mitigazione dell'effetto di amplificazione dell'intensità di \vec{B} .

Quando si trattano fenomeni di accoppiamento magnetico il materiale ferromagnetico produce un ulteriore effetto rilevante, che può essere distinto in modo sufficientemente chiaro. Ad esso diamo qui il nome di *canalizzazione* delle linee di campo di induzione magnetica. Mostriamo questo effetto facendo riferimento a un argomento ampiamente discusso in ogni testo di Fisica Generale e spesso accompagnato dal nome (molto misleading) di "rifrazione" del campo magnetico. Le equazioni di Maxwell che definiscono i campi \vec{B} e \vec{H} sono

$$\vec{\nabla} \cdot \vec{B} = 0 \quad (13)$$

$$\vec{\nabla} \times \vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t}, \quad (14)$$

con \vec{J} densità di corrente elettrica e \vec{D} campo elettrico nella materia (campo "di spostamento"). Supponendo di non avere densità di correnti (di cariche "libere") che fluiscono nel materiale ferromagnetico o sulla sua superficie, si può porre $\vec{J} = 0$.

Immaginiamo allora di avere un'interfaccia che divide due regioni di spazio, interne ed esterne a un materiale ferromagnetico, e indichiamo con \vec{B}_{int} e \vec{B}_{ext} i campi

di induzione magnetica nelle due regioni di spazio. Per costruire un problemino di magnetostatica, supponiamo di conoscere \vec{B}_{ext} , indicando con θ_{ext} l'angolo che tale vettore forma con la normale all'interfaccia in un certo punto. Usando una simbologia che fa riferimento alle componenti parallele e ortogonali all'interfaccia, si ha $B_{//ext} = B_{ext} \sin \theta_{ext}$ e $B_{\perp ext} = B_{ext} \cos \theta_{ext}$.

Le Eqs. 13,14, integrate opportunamente in un intorno del punto considerato, la prima su un volume delimitato da una superficie chiusa Σ e la seconda su una superficie delimitata da una linea chiusa γ , danno luogo a

$$\int_{\Sigma} \vec{B} \cdot \hat{n} d\Sigma = 0 \quad (15)$$

$$\oint_{\gamma} \vec{H} \cdot d\vec{l} = \frac{d\Phi_S(\vec{D})}{dt}, \quad (16)$$

dove \hat{n} è il versore uscente dalla superficie di integrazione Σ in ogni suo punto, $d\vec{l}$ è elemento di γ , $\Phi_S(\vec{D})$ rappresenta il flusso di \vec{D} attraverso una superficie delimitata dalla linea chiusa γ .

Poiché siamo interessati a verificare il comportamento dei campi al passaggio per l'interfaccia, possiamo scegliere in maniera opportuna Σ e γ , per esempio considerando un barattolo con asse parallelo alla normale dell'interfaccia e altezza molto piccola (infinitesima) per Σ e un rettangolo con altezza parallela alla normale all'interfaccia, e anch'essa resa molto piccola (infinitesima), per γ . In queste condizioni la superficie S su cui viene calcolato $\Phi_S(\vec{D})$ tende a zero, per cui al primo ordine si può approssimare a zero il secondo membro dell'Eq. 16.

Le Eqs. 15,16 forniscono delle *condizioni di continuità* sulle componenti dei campi che, con la nostra simbologia, recitano

$$B_{\perp ext} = B_{\perp int} \quad (17)$$

$$H_{//ext} = H_{//int}. \quad (18)$$

Poiché supponiamo valida l'Eq. 12, l'Eq. 18 si può riscrivere come

$$B_{//ext} = \frac{B_{//int}}{\mu_r}, \quad (19)$$

dove, coerentemente con la descrizione del problema, abbiamo posto $\mu_{r,ext} = 1$ (materiale non ferromagnetico, per esempio l'aria) e $\mu_{r,int} = \mu_r$.

Per la geometria si ha, con ovvia definizione dei simboli,

$$\tan \theta_{int} = \frac{B_{//int}}{B_{\perp int}} = \mu_r \frac{B_{//ext}}{B_{\perp ext}} = \mu_r \tan \theta_{ext}. \quad (20)$$

Allora, se $\mu_r >> 1$ come stiamo imponendo, per qualsiasi valore di θ_{ext} si ha $\theta_{int} \rightarrow \pi/2$, cioè *le linee del campo di induzione magnetica all'interno di un materiale ferromagnetico tendono a disporsi parallelamente all'interfaccia*. In altre parole, un materiale ferromagnetico di una certa forma costituisce una sorta di tubo di flusso per \vec{B} , per cui

le linee del campo sono *canalizzate* all'interno della forma stessa. Questo effetto “geometrico” di canalizzazione può sicuramente essere sfruttato per “trasportare” le linee di campo da una bobina all'altra e quindi per aumentare il coefficiente di accoppiamento magnetico k . Tuttavia, oltre a implicare i fenomeni di dissipazione e di schermatura del campo ricordati prima, come approfondiremo meglio in futuro la canalizzazione attuata da un tubo di materiale ferromagnetico che *non si richiude su se stesso* a formare un circuito è generalmente insufficiente per ottenere $k \simeq 1$.

III. PRIMARIO E SECONDARIO IN DIVERSE CONFIGURAZIONI

Limitandoci alle condizioni delle Eqs. 8,9 (due induttori accoppiati magneticamente, segnali sinusoidali, presenza di soli elementi resistivi nei circuiti oltre a auto e mutua induzione), ci proponiamo ora di esplorare diverse configurazioni realizzabili sperimentalmente. Esse sono significative per apprezzare gli effetti della mutua induzione e inoltre costituiscono una valida introduzione allo studio del trasformatore, un dispositivo elettrotecnico di cui tratteremo nel futuro.

Noteate che negli esperimenti svolti in laboratorio abbiamo praticamente avuto accesso solo ad alcune di queste configurazioni (in sostanza, secondario cortocircuitato e induttori in serie). Tuttavia è utile e opportuno fare un esame un po' più ampio.

A. Secondario aperto

In questa prima configurazione immaginiamo che un generatore di forme d'onda sinusoidali fornisca la d.d.p. descritta dal fasore $V_{\omega 1}$ al circuito primario (per comodità possiamo supporre il generatore ideale) e che il secondario sia “aperto”, cioè non supporti passaggio di corrente ($I_{\omega 2} = 0$). Questa situazione assomiglia a quella che si può verificare sperimentalmente collegando al secondario uno strumento di misura per la lettura della d.d.p. (oscilloscopio, multmetro) dotato di elevata resistenza, o impedenza, di ingresso.

Le Eqs. 8,9 diventano

$$V_{\omega 1} = (R_1 + j\omega L_1)I_{\omega 1} \quad (21)$$

$$V_{\omega 2} = j\omega M I_{\omega 1}, \quad (22)$$

da cui è possibile esprimere la *funzione di trasferimento* T_{open} tale che $V_{\omega 2} = T_{open}V_{\omega 1}$:

$$T_{open} = \frac{j\omega M}{R_1 + j\omega L_1}. \quad (23)$$

Nell'ipotesi $\omega L_1 \gg R_1$, spesso verificata nelle applicazioni pratiche, si ha in prima approssimazione $T_{open} \simeq M/L_1 = k\sqrt{L_2/L_1}$, dove abbiamo impiegato la definizione di Eq. 10. In queste condizioni la funzione di trasferimento è reale e non c'è sfasamento tra $V_{\omega 2}$ e $V_{\omega 1}$ (ovvero

lo sfasamento è trascurabile). Il rapporto tra le ampiezze, o ampiezze picco-picco, dei due segnali, che, usando un termine gergale, individua il *rapporto di trasformazione in tensione*, è $T_V = V_2/V_1 = k\sqrt{L_2/L_1}$. Progettando in maniera opportuna il sistema dei due induttori accoppiati, è possibile modificare a volontà l'ampiezza, o ampiezza picco-picco, della d.d.p. al secondario rispetto a quella nel primario, sia diminuendola che aumentandola.

B. Secondario cortocircuitato

In questa configurazione supponiamo di cortocircuittare il secondario, cioè di imporre $V_{\omega 2} = 0$. Le Eqs. 8,9 diventano:

$$V_{\omega 1} = (R_1 + j\omega L_1)I_{\omega 1} + j\omega M I_{\omega 2} \quad (24)$$

$$0 = (R_2 + j\omega L_2)I_{\omega 2} + j\omega M I_{\omega 1}. \quad (25)$$

Da Eq. 25 si ricava

$$I_{\omega 2} = -\frac{j\omega M}{R_2 + j\omega L_2} I_{\omega 1}, \quad (26)$$

che, sostituita in Eq. 24, conduce a

$$V_{\omega 1} = \left[R_1 + j\omega L_1 + j\omega M \frac{-j\omega M}{R_2 + j\omega L_2} \right] I_{\omega 1} = \quad (27)$$

$$= \left[R_1 + \frac{j\omega L_1 R_2 - \omega^2 (L_1 L_2 - M^2)}{R_2 + j\omega L_2} \right] I_{\omega 1} = \quad (28)$$

$$= \left[R_1 + \frac{j\omega L_1 R_2 - \omega^2 L_1 L_2 (1 - k^2)}{R_2 + j\omega L_2} \right] I_{\omega 1}. \quad (29)$$

Nell'ipotesi $\omega L_2 \gg R_2$, spesso verificata nelle applicazioni pratiche (attenzione: la condizione si riferisce al secondario, e si può sicuramente avere $L_2 \neq L_1$), si ha in prima approssimazione

$$V_{\omega 1} = \left[R_1 + R_2 \frac{L_1}{L_2} + j\omega L_1 (1 - k^2) \right] I_{\omega 1}. \quad (30)$$

Nell'ulteriore ipotesi di accoppiamento magnetico completo, $k \simeq 1$, realizzata nel trasformatore *ideale*, la relazione tra intensità di corrente $I_{\omega 1}$ e d.d.p. $V_{\omega 1}$ è *reale*, per cui, nonostante la presenza di impedenze reattive (immaginarie), al primario lo sfasamento tra tensione e corrente è trascurabile. In particolare il generatore $V_{\omega 1}$ vede un carico *resistivo* di valore pari alla somma di R_1 e $R_2 L_1 / L_2$.

Torniamo ora all'Eq. 26, che, in pratica, stabilisce la funzione di trasferimento T_{short} delle intensità di corrente nei due circuiti, essendo tale che $I_{\omega 2} = T_{short} I_{\omega 1}$. Nell'ipotesi, già utilizzata, $\omega L_2 \gg R_2$, la T_{short} diventa approssimativamente *reale*, $T_{short} \simeq -M/L_2 = -k\sqrt{L_1/L_2}$, e stabilisce uno sfasamento di π rad tra segnali di corrente al secondario e al primario (che questo sfasamento possa effettivamente essere osservato dipende dalla conoscenza dei versi di avvolgimento delle spire degli induttori!). Il rapporto tra le ampiezze,

o ampiezze picco-picco, di questi segnali, generalmente detto *rapporto di trasformazione in corrente*, è quindi $T_A = k\sqrt{L_1/L_2}$. Il rapporto dipende quindi dal progetto del sistema, cioè dai valori dei coefficienti di autoinduzione e di accoppiamento magnetico.

C. Serie e anti-serie di induttori

Supponiamo ora di collegare in serie i due induttori, realizzando un circuito al cui ingresso colleghiamo un generatore di forme d'onda che produce la d.d.p. sinusoidale descritta dal fasore V_ω . Nel circuito, in cui immaginiamo sia presente anche una resistenza complessiva R , eventualmente comprensiva delle resistenze interne dei due induttori, scorre una corrente la cui intensità è descritta dal fasore I_ω . A causa dell'accoppiamento magnetico, la variazione di corrente su uno dei due induttori induce una d.d.p. ai capi dell'altro. L'equazione del circuito nel dominio delle frequenze si scrive

$$V_\omega = [R + j\omega(L_1 \pm M) + j\omega(L_2 \pm M)] I_\omega \quad (31)$$

dove abbiamo distinto gli effetti della mutua induzione per gli induttori 1 e 2.

I segni \pm in Eq. 31 riflettono la possibilità che le due bobine possano essere realizzate avvolgendo il filo nello stesso verso, o in verso opposto, rispetto a un riferimento comune, per esempio l'asse delle bobine. Spesso negli schemi circuitali il verso di circolazione delle correnti nei due avvolgimenti può essere distinto grazie alla presenza di un pallino, o un qualche altro simbolo, disegnato accanto a uno dei fili dell'avvolgimento a rappresentarne "l'ingresso". Come già discusso, questa possibilità si riflette in d.d.p. indotte di segno opposto, ovviamente per entrambi gli induttori. L'Eq. 31 mostra che l'*induttanza equivalente* (o efficace, o effettiva, o totale) della serie è

$$L_{eq,serie} = L_1 + L_2 \pm 2M. \quad (32)$$

L'accoppiamento magnetico determina un termine aggiuntivo, di segno opportuno, che modifica il carattere additivo delle impedenze in serie, e dunque delle sue componenti reattive, cioè delle induttanze coinvolte.

D. Parallelo e anti-parallelo di induttori

Supponiamo qui di avere un collegamento in parallelo tra i due induttori. Poniamo trascurabili tutte le resisten-

ze per semplicità e chiamiamo V_ω il fasore che descrive la d.d.p. sinusoidale applicata al parallelo e $I_{\omega 1,2}$ i fasori che descrivono le intensità di corrente nei due induttori. Si ha:

$$V_\omega = j\omega(L_1 I_{\omega 1} \pm M I_{\omega 2}) = \quad (33)$$

$$= V_\omega = j\omega(L_2 I_{\omega 2} \pm M I_{\omega 1}), \quad (34)$$

da cui

$$I_{\omega 1} = I_{\omega 2} \frac{L_2 \mp M}{L_1 \mp M}, \quad (35)$$

dove, al solito, i segni \pm e \mp dipendono dai versi di circolazione della corrente nei due avvolgimenti. Ricavando $I_{\omega 1}$ da Eq. 33 e sostituendolo in Eq. 34 si ottiene

$$I_{\omega 2} = \frac{V_\omega}{j\omega} \frac{1 \mp M/L_1}{L_2 - M^2/L_1} = \frac{V_\omega}{j\omega} \frac{L_1 \mp M}{L_1 L_2 - M^2}. \quad (36)$$

D'altra parte, per la conservazione della carica deve essere

$$I_\omega = I_{\omega 1} + I_{\omega 2} = \left(\frac{L_2 \mp M}{L_1 \mp M} + 1 \right) I_{\omega 2} = \quad (37)$$

$$= \frac{L_2 \mp M + L_1 \mp M}{L_1 \mp M} \frac{V_\omega}{j\omega} \frac{L_1 \mp M}{L_1 L_2 - M^2} = \quad (38)$$

$$= \frac{V_\omega}{j\omega} \frac{L_1 + L_2 \mp 2M}{L_1 L_2 - M^2}. \quad (39)$$

L'Eq. 39 stabilisce che l'*induttanza equivalente* del parallelo è

$$\begin{aligned} L_{eq,para} &= \frac{1}{j\omega} \frac{V_\omega}{I_\omega} = \frac{L_1 L_2 - M^2}{L_1 + L_2 \mp 2M} = \\ &= L_1 L_2 \frac{1 - k^2}{L_1 + L_2 \mp k\sqrt{L_1 L_2}}, \end{aligned} \quad (40)$$

dove, al solito, il segno \mp si riferisce alla concordanza (segno "-") o discordanza (segno "+") dei versi di percorrenza della corrente nei due avvolgimenti e k è il coefficiente di accoppiamento magnetico. Nel caso $M \simeq 0$ si ritrova, ovviamente, la regola del parallelo di induttanze non accoppiate: l'Eq. 40 mostra quindi l'effetto dell'accoppiamento magnetico nel parallelo di induttori.

Onde elettromagnetiche, ottica, polarizzazione

francesco.fuso@unipi.it; <http://www.df.unipi.it/~fuso/dida>

(Dated: version 5 - FF, 11 maggio 2017)

In questa nota vengono richiamati alcuni concetti e alcuni formalismi utili per trattare semplici situazioni in cui sono coinvolte onde elettromagnetiche, in particolare nell'ambito dell'ottica. Queste situazioni sono parte di quelle che si incontrano nelle esperienze pratiche di laboratorio, in particolare quelle che riguardano la polarizzazione, la sua misura e la sua manipolazione. L'argomento delle onde elettromagnetiche è estremamente vasto e ricco di spunti concettuali che in queste note saranno bellamente ignorati. L'enfasi viene posta, piuttosto, su alcune questioni di terminologia, nomenclatura e semplice matematica, utili per avere un background sufficientemente ampio per la comprensione delle esperienze pratiche, oltre che per avere un quadro sufficientemente generale dell'ottica di polarizzazione.

I. EQUAZIONE E FUNZIONE D'ONDA

Una *funzione d'onda* rappresenta in generale l'andamento nello spazio e nel tempo di una qualche “perturbazione”. Per esempio, una funzione d'onda può descrivere l'andamento di campo elettrico e magnetico nello spazio e nel tempo. Dal punto di vista matematico, la funzione d'onda è soluzione di una *equazione d'onda*, costruita sulla base di definizioni e equazioni specifiche per le grandezze fisiche e il sistema che si stanno considerando.

L'equazione d'onda per le onde elettromagnetiche in un materiale dielettrico, *in assenza di correnti e cariche libere*, si costruisce a partire dalle equazioni di Maxwell per il rotore dei campi. Si ha infatti

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (1)$$

$$\vec{\nabla} \times \vec{B} = \mu \epsilon \frac{\partial \vec{E}}{\partial t}, \quad (2)$$

dove la seconda equazione deriva da $\vec{\nabla} \times \vec{H} = \partial \vec{D} / \partial t$ con l'assunzione, valida per materiali isotropi, omogenei e “lineari”, $\vec{B} = \mu_0 \mu_r \vec{H} = \mu \vec{H}$ e $\vec{D} = \epsilon_0 \epsilon_r \vec{E} = \epsilon \vec{E}$. Per i fenomeni che intendiamo discutere qui, che riguardano materiali di tipo ordinario (non nanostrutturati, non “metamateriali”), possiamo sicuramente porre $\mu = \mu_0$, ovvero $\mu_r = 1$: infatti μ_r è sensibilmente diverso da uno solo per i ferromagneti, che sono generalmente conduttori opachi, e che quindi non permettono il passaggio della radiazione elettromagnetica di interesse per l'ottica.

Facendo il rotore dei due membri delle due equazioni di Maxwell e usando un minimo di algebra si ottiene facilmente l'equazione d'onda per il campo elettrico

$$\nabla^2 \vec{E} = \mu_0 \epsilon \frac{\partial^2 \vec{E}}{\partial t^2}; \quad (3)$$

un'equazione formalmente identica si ottiene anche per il campo di induzione magnetica. Considerato ciò, è lecito focalizzarsi sulla soluzione per il campo elettrico, essendo sempre possibile, come vedremo dopo, ricavare l'andamento del campo di induzione magnetica.

È molto interessante notare che il termine $\mu_0 \epsilon$ che compare al secondo membro dell'Eq. 3 deve avere, per sem-

plici ragioni dimensionali, le dimensioni di una velocità alla meno due. Quindi è possibile individuare una velocità (detta *velocità di fase*, e sarà l'unica velocità di cui tratteremo qui, non essendo interessati a fenomeni “dispersivi”) $v = 1/\sqrt{\mu_0 \epsilon}$. Tenendo conto che nel vuoto, dove $\epsilon = \epsilon_0$, fino a prova contraria tale velocità è la costante $c \approx 3 \times 10^8$ m/s, in un materiale dotato di permittività dielettrica relativa $\epsilon_r > 1$ si ha $v = c/\sqrt{\epsilon_r} = c/n$, dove è stato introdotto l'*indice di rifrazione (reale)* del mezzo considerato, $n = \sqrt{\epsilon_r}$ (con $\epsilon_r \geq 1$ reale). Questa definizione è in accordo con quello che tutti diamo per buona, cioè che una perturbazione (elettromagnetica, o di qualsiasi altro genere) viaggia a una velocità inferiore o al massimo uguale alla velocità della luce.

A. Onda piana, monocromatica, progressiva

Una soluzione dell'Eq. 3 è data dalla funzione d'onda

$$\vec{E} = \{E_{0+} \exp[i(\vec{k} \cdot \vec{r} - \omega t)] + E_{0-} \exp[i(-\vec{k} \cdot \vec{r} - \omega t)]\} \hat{e}. \quad (4)$$

Per motivi legati all'accresciuta semplicità matematica che così viene offerta, la funzione è scritta in forma complessa: come si verifica sempre quando si usano grandezze complesse, il campo elettrico dell'onda elettromagnetica è dato dalla parte reale dell'espressione.

Nella funzione di Eq. 4 compaiono due termini, sommati tra loro attraverso i pesi E_{0+} e E_{0-} , che devono avere le dimensioni di campi elettrici. Grazie alla linearità delle equazioni di cui facciamo uso, sicuramente non si perde in generalità se ci si limita a considerare solo uno dei due termini, per cui considereremo come funzione d'onda “modello” per il campo elettrico la

$$\vec{E} = E_0 \exp[i(\vec{k} \cdot \vec{r} - \omega t)] \hat{e}, \quad (5)$$

che, per i motivi che illustreremo tra breve, rappresenta un'*onda piana, monocromatica, progressiva* e armonica. Prima di proseguire, scriviamo anche la soluzione dell'equazione d'onda, ovvero la funzione d'onda, per il campo di induzione magnetica: come già osservato, essa deve formalmente essere simile a quella per il campo elettrico,

e quindi deve avere l'espressione

$$\vec{B} = B_0 \exp[i(\vec{k} \cdot \vec{r} - \omega t)]\hat{b}. \quad (6)$$

Osservate che nelle espressioni compaiono diversi versori (\hat{e} e \hat{b}) a indicare le direzioni dei due campi, che, come stabilito dalle equazioni di Maxwell e riassunto dalle proprietà dell'operatore rotore, non possono essere le stesse. Infatti, affinché le equazioni di Maxwell siano soddisfatte, deve essere, come è facile verificare,

$$\vec{B} = \frac{\hat{k} \times \vec{E}}{v}, \quad (7)$$

ovvero le direzioni di $\vec{k}, \vec{E}, \vec{B}$ formano una *terna ortogonale destrorsa*, per cui l'onda si dice *trasversale*. In altre parole, nelle onde di nostro interesse campo elettrico e magnetico sono ortogonali fra loro, ed entrambi giacciono su un piano che è ortogonale alla direzione di \vec{k} . Inoltre dall'Eq. 7 si deduce che le ampiezze dei campi sono legate tra loro dalla relazione

$$B_0 = \frac{E_0}{v} = \frac{E_0}{c} n. \quad (8)$$

Ricordando la definizione di *fronte d'onda* come luogo dei punti in cui, a un dato istante, la perturbazione (i campi!) hanno un dato valore, si vede subito che le funzioni che abbiamo scelto come soluzione dell'equazione d'onda sono:

- *piane*, essendo i fronti d'onda dei piani ortogonali alla direzione del vettore d'onda \vec{k} ;
- *progressive*, perché i fronti d'onda si muovono alla velocità di fase v nella direzione positiva di \vec{k} , come potete facilmente verificare calcolando dove "va a finire" un certo fronte d'onda dopo un certo intervallo di tempo;
- *monocromatiche*, dato che l'andamento temporale è stabilito da un'unica pulsazione ω ;

Inoltre si usa l'aggettivo armoniche per indicare che, in una posizione fissata, l'intensità della perturbazione (dei campi) è descritta nel dominio del tempo da funzioni tipo seno o coseno.

B. Trasporto di energia e intensità dell'onda

La rilevanza che le onde elettromagnetiche hanno in moltissimi fenomeni risiede principalmente nel fatto che esse sono associate a trasporto di energia, ovvero di potenza. Nelle nostre esperienze pratiche, questo trasporto di energia viene analizzato usando un fotorivelatore, come ad esempio un fotodiode. Un dispositivo di questo tipo permette di misurare il *valore medio nel tempo* della potenza dell'onda integrata su una certa superficie (la minore tra sezione del fascio luminoso che si utilizza e la

sezione dell'area sensibile del dispositivo). D'altra parte, come mostrato nella Fig. 1, che commenteremo fra breve, le onde elettromagnetiche rilevanti in ottica hanno frequenze, o pulsazioni, che non possono essere apprezzate da nessuno strumento puramente elettronico (per intenderci, un oscilloscopio), per cui è evidente che quello che viene valutato e misurato è il valore medio nel tempo.

L'energia trasportata da un'onda elettromagnetica per unità di tempo e di superficie è data dal vettore di Poynting, $\vec{S} = \vec{E} \times \vec{H} = \vec{E} \times \vec{B}/\mu_0$. Per l'ortogonalità tra le direzioni che abbiamo stabilito sopra, \vec{S} ha la direzione di \vec{k} per un'onda progressiva. Dunque la direzione di \vec{k} , detta *direzione di propagazione*, è quella lungo cui si svolge il trasferimento di energia. Il calcolo del valore medio nel tempo può essere fatto con diverse tecniche. La più elegante è la seguente:

$$\langle \vec{S} \rangle = \frac{1}{2} \operatorname{Re}\{\vec{E} \times \vec{H}^*\}. \quad (9)$$

Eseguendo il calcolo per il *modulo* di questa quantità, tenendo conto della relazione tra le ampiezze dei campi data in precedenza, si trova facilmente

$$I = |\langle \vec{S} \rangle| = \frac{1}{2} v \epsilon E_0^2. \quad (10)$$

L'espressione $\epsilon E_0^2/2$ rappresenta la densità di energia media associata al campo elettromagnetico, per cui la grandezza sopra espressa rappresenta il flusso di potenza su una superficie unitaria ortogonale alla direzione di \vec{k} , e ha le dimensioni di una potenza diviso per una superficie (una buona unità di misura potrebbe essere W/m^2 , o W/cm^2 , come spesso in uso in ottica laser). A essa si dà il nome (gergale) di *intensità* dell'onda.

C. Semplificazione, nomenclatura pratica e ordini di grandezza

Per semplificare la notazione, consideriamo d'ora in avanti $\vec{k} // \hat{z}$, cioè assumiamo che l'onda piana si propaghi lungo la direzione dell'asse Z di un sistema di riferimento cartesiano. Di conseguenza, campo elettrico e magnetico oscillano in fase, ortogonali fra loro, sul piano XY e la funzione d'onda di Eq. 5 assume la forma

$$\vec{E} = E_0 \exp[i(kz - \omega t)]\hat{e}, \quad (11)$$

con \hat{e} appartenente al piano XY .

Si parla comunemente di onde elettromagnetiche a proposito di fenomeni che coinvolgono range estremamente ampi dei parametri k e ω che compaiono nella funzione d'onda. È possibile fare un minimo di classificazione basandosi soprattutto sulle applicazioni delle onde coinvolte. Un esempio è fornito dalla Fig. 1 (tratta da wikipedia), dove la classificazione è fatta sulla base del valore della *lunghezza d'onda* λ , su cui torneremo fra breve.

Nell'ottica (tradizionale) si lavora in genere con radiazione elettromagnetica visibile (o nel range compreso tra

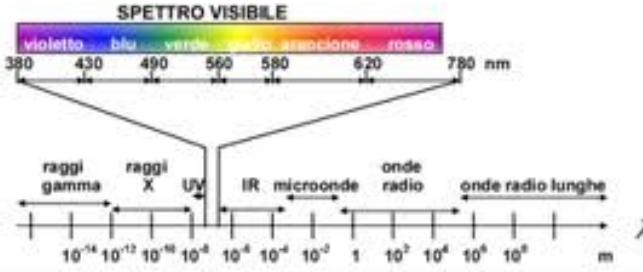


Figura 1. Esempio di classificazione delle onde elettromagnetiche sulla base della lunghezza d'onda λ e rappresentazione a colori dello spettro visibile.

vicino ultravioletto - UV - e vicino infrarosso - IR) a cui convenzionalmente corrisponde una lunghezza d'onda di alcune centinaia di nanometri (per esempio, $\lambda = 380\text{--}780$ nm, circa, per il visibile). In questo modo ci si restringe di fatto a una piccola fettina del cosiddetto spettro della radiazione elettromagnetica.

Può essere utile disegnare il grafico della parte reale della funzione d'onda di Eq. 11 in funzione della coordinata z (supponendo fissato il tempo, per esempio a $t = 0$) o in funzione del tempo t (supponendo fissata la posizione, per esempio in $z = 0$), come rappresentato in Fig. 2. Questa operazione permette di apprezzare gli ordini di grandezza delle scale spaziali e temporali su cui varia il campo. Si vede facilmente come la lunghezza d'onda λ sia legata al numero d'onda $k = |\vec{k}|$, alla frequenza $\nu = 2\pi/\omega$ e alla velocità di fase v (che qui supponiamo pari a c , cioè l'onda *si propaga nel vuoto*, in questo caso), dalle relazioni

$$k = \frac{2\pi}{\lambda} \quad (12)$$

$$\nu = \frac{c}{\lambda}. \quad (13)$$

Gli ordini di grandezza (e le unità di misura) per numero d'onda e frequenza rilevanti in ottica sono quindi: $k \sim 10^4 \text{ cm}^{-1}$ (notate l'unità di misura, che è tipica nell'ottica), $\nu \sim (4\text{--}9) \times 10^{14} \text{ Hz}$ (notate l'esponentone), che, non per un caso, è dello stesso ordine di grandezza della frequenza di rotazione dell'elettrone attorno al nucleo in un modello atomico classico (planetario).

D. Energia della radiazione

Un altro parametro caratteristico della radiazione, che ha grosse implicazioni in tantissimi ambiti, riguarda l'*energia* della radiazione stessa. L'argomento che permette di valutare questo parametro in modo diretto ha a che fare con aspetti non classici della fisica, che avrete ampio modo di approfondire andando avanti con i vostri studi. Tuttavia, già in questa sede può far comodo notare che, in parallelo alla descrizione *ondulatoria* della radiazione elettromagnetica, è possibile utilizzare una descrizione *corpuscolare*, in conseguenza della quale, per

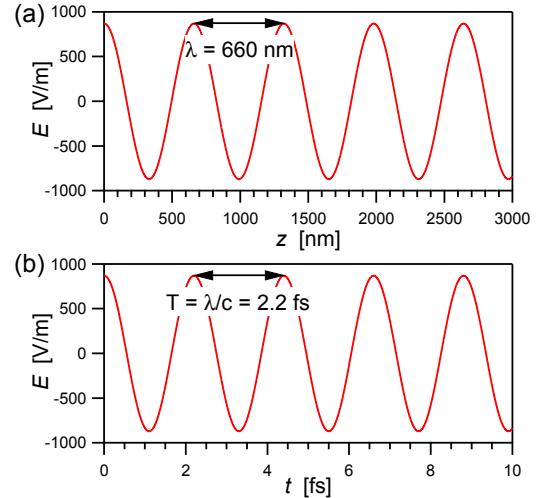


Figura 2. Grafico della parte reale della funzione d'onda per il campo elettrico (onda progressiva diretta lungo \hat{z}) in funzione della coordinata z (a) e del tempo t (b). Nel grafico (a) si è supposto $t = 0$ e imposto che il campo fosse massimo in $z = 0$; nel grafico (b) si è supposto $z = 0$ e imposto che il campo fosse massimo in $t = 0$. Lunghezza d'onda λ e periodo $T = 2\pi/\omega$ sono indicati nei grafici. Per fissare le scale orizzontali dei grafici si è scelta $\lambda = 660 \text{ nm}$ (un bel colore rosso intenso). Per fissare la scala verticale si è supposto che la radiazione avesse la potenza $P = 1 \text{ mW}$, che la propagazione avvenisse nel vuoto e che si fosse in presenza di un fascio di radiazione con distribuzione spaziale di intensità omogenea su un'area di 1 mm^2 (tutto questo consente di determinare l'ampiezza del campo elettrico dell'onda): notate bene i valori numerici e le unità di misura delle scale.

intenderci, un fascio di luce può essere visto come un flusso di particelle con proprietà molto specifiche (senza massa, senza carica, ma dotate di quantità di moto, momento angolare e, appunto, energia), che si chiamano *fotoni*. L'energia trasportata da un singolo fotone dipende solo dalla frequenza della radiazione attraverso la semplice relazione

$$E_{phot} = h\nu, \quad (14)$$

con h costante fondamentale della meccanica quantistica, chiamata *costante di Planck* ($h \approx 6.6 \times 10^{-34} \text{ J s}$). Nel vuoto, la relazione appena scritta dà luogo a un legame di proporzionalità inversa con la lunghezza d'onda: $E_{phot} = hc/\lambda$. Il prodotto tra le costanti fondamentali h e c può essere determinato numericamente. Alla fine, si ottiene una formulina che siete invitati a tenere sempre bene in mente:

$$E_{phot} [\text{eV}] \approx \frac{1240}{\lambda [\text{nm}]} . \quad (15)$$

Dunque il fotone della radiazione visibile ha un'energia dell'ordine di 1.5 – 3 eV (la scala di energie in eV è quella più semplice da usare quando si vogliono descrivere le

proprietà della materia, incluse quelle che riguardano la sua interazione con la luce).

Se ricordate che $1 \text{ eV} \cong 1.6 \times 10^{-19} \text{ J}$, potete facilmente rendervi conto che un singolo fotone porta un'energia piccolissima. Però di fotoni potete (spesso) averne tantissimi: per esempio, un piccolo puntatore laser di potenza 1 mW che emette nel rosso ($\lambda \approx 650 \text{ nm}$) è in grado di produrre qualcosa dell'ordine di 10^{16} fotoni al secondo.

E. Limitazioni e caveat

Ci sono un paio di osservazioni che è opportuno citare a questo punto. La prima riguarda il modo con cui abbiamo ottenuto l'equazione d'onda e le varie semplificazioni, o approssimazioni, che abbiamo, anche implicitamente, usato. In primo luogo, la maggior parte dei materiali di interesse per l'ottica, e più in generale per la fisica, sono *dispersivi*, cioè in questi materiali l'indice di rifrazione n dipende dalla lunghezza d'onda, o frequenza, della radiazione. Di per sé questa precisazione non modifica la matematica che abbiamo sviluppato, però ricordatevene. Fra le tante conseguenze, nel caso di materiali dispersivi, accanto alla velocità di fase $v = \omega/k$, si definisce un'altra velocità, detta *velocità di gruppo*, $v_g = \partial\omega/\partial k$, che serve proprio a tenere conto di come l'onda si propaga in un mezzo il cui indice di rifrazione cambia con la frequenza dell'onda stessa.

Inoltre, come già accennato, in questa nota ci limitiamo a considerare situazioni fisiche "ordinarie" per l'ottica, in cui, per esempio, $\mu_r \approx 1$, $\epsilon_r \geq 1$ reale, e $n = \sqrt{\epsilon_r}$ reale. La tecnologia attuale è certamente in grado di creare dei sistemi artificiali, micro- o nanostrutturati, il cui comportamento macroscopico può essere descritto (abbastanza) facilmente ipotizzando che le grandezze che abbiamo citato abbiano valori "non convenzionali", o siano immaginarie. Avrete probabilmente sentito parlare di cristalli fotonici, metamateriali, superfici stealth, plasmoni, tutti ambiti in cui è spesso conveniente definire in maniera opportuna, e diversa da quella convenzionale, le grandezze di interesse. Bene, tutto questo qui non lo consideriamo.

Infine, è necessario chiarire che l'uso dell'onda piana, progressiva e monocromatica come funzione "modello" dell'ottica contiene alcune criticità, che possono essere individuate (ma non sanate) facilmente. L'onda piana porta con sé un'idea di "infinito" che non sempre suona realistica: infatti i fronti d'onda (piani) che essa prevede hanno virtualmente un'estensione infinita (non c'è alcuna dipendenza da x o y nella funzione di Eq. 11) e il carattere puramente monocromatico implica, come sapete o saprete, che l'onda è stata accesa in un istante infinitamente precedente e che verrà spenta in un istante infinitamente successivo a quello dell'osservazione. Tuttavia, specie adottando alcune ulteriori tecniche di rappresentazione ("modi trasversali", "pacchetti d'onda"), si può convivere con queste difficoltà, soprattutto perché la grande semplicità che offre la trattazione di problemi con onde piane supera (spesso) le inesattezze che essa comporta.

II. POLARIZZAZIONE

Generalmente, lo studio, ovvero la misura e la manipolazione, della *polarizzazione* riguarda l'analisi del versore \hat{e} che indica la *direzione del campo elettrico* (per le onde che qui consideriamo, quella del campo magnetico è sempre ortogonale a questa e alla direzione di propagazione). Supponendo un'onda piana che si propaga lungo Z , sappiamo che \hat{e} deve appartenere al piano XY , ma in questo modo non ne fissiamo la direzione.

La polarizzazione si dice *lineare* se \hat{e} mantiene costante nel tempo la sua direzione, cioè se il campo elettrico oscilla *sempre* lungo la *stessa* direzione cartesiana. Questo si ottiene, ad esempio, scrivendo $\hat{e} = a\hat{x} + b\hat{y}$, con a, b costanti reali opportunamente normalizzate. La direzione di \hat{e} si mantiene allora costante nel tempo, essendo individuata ad esempio dall'angolo ϕ rispetto alla direzione X : $\tan \phi = b/a$.

La polarizzazione si dice *circolare* quando, invece, il vettore \vec{E} ruota con velocità angolare ω sul piano XY . Questo si ottiene, ad esempio, quando $\hat{e} = (\hat{x} \pm i\hat{y})/\sqrt{2}$ (notate il fattore di normalizzazione), cioè quando le componenti E_x e E_y del campo sono *sfasate fra loro* di $\pm\pi/2$. In queste condizioni possiamo infatti scrivere

$$\vec{E} = \frac{E_0}{\sqrt{2}} \{(\exp[i(kz - \omega t)])\hat{x} + (\exp[i(kz - \omega t \pm \pi/2)])\hat{y}\}, \quad (16)$$

e notare che il termine aggiunto nell'argomento dell'esponentiale complesso per la componente Y , $\exp(\pm i\pi/2)$, equivale a moltiplicare per $\pm i$ la componente stessa. Se estraiamo la parte reale di questa funzione d'onda possiamo facilmente osservare che la "punta" del vettore \vec{E} compie una traiettoria circolare sul piano XY , con velocità angolare ω .

Il segno dello sfasamento denota due tipi di polarizzazione circolare, la *circolare sinistra* e la *circolare destra*, che spesso si indicano con $\hat{\sigma}_-$ e $\hat{\sigma}_+$, a cui corrispondono due sensi di rotazione. In genere, la convenzione si riferisce a un osservatore verso il cui occhio si dirige la radiazione; sinistra e destra hanno allora a che fare con rotazione rispettivamente antioraria o oraria. È interessante osservare che, come \hat{x} e \hat{y} , anche $\hat{\sigma}_-$ e $\hat{\sigma}_+$ sono *basi ortonormali* per descrivere il vettore \vec{E} nel piano XY . In altre parole, qualsiasi direzione di polarizzazione lineare può essere rappresentata con opportune combinazioni lineari delle polarizzazioni circolari destra e sinistra.

La Fig. 3 illustra schematicamente la polarizzazione lineare e quella circolare sul piano XY .

Infine, la polarizzazione si dice *ellittica* quando lo sfasamento tra le componenti X e Y di \vec{E} è diverso in modulo da $\pi/2$ (o suoi multipli dispari), ovvero, che è la stessa cosa, quando $\hat{e} = a\hat{x} + b\hat{y}$, con a, b costanti *complesse* opportunamente normalizzate: è abbastanza facile convincersi che, per polarizzazione ellittica, la "punta" di \vec{E} descrive un'ellisse nel piano XY , e che la polarizzazione circolare può essere considerata un caso particolare di polarizzazione ellittica.

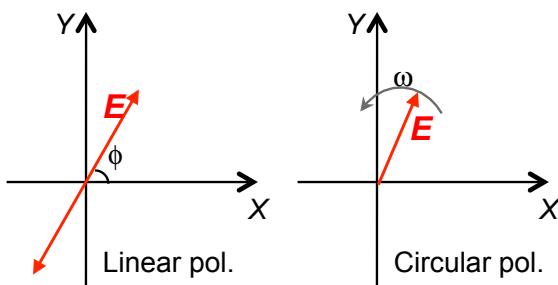


Figura 3. Illustrazione schematica sul piano XY della polarizzazione lineare (che forma un angolo ϕ rispetto all'asse X) e circolare (che ruota in senso antiorario, dunque di tipo $\hat{\sigma}_-$).

A. Significato e rilevanza della polarizzazione

Avere una polarizzazione lineare implica, di fatto, essere in grado di individuare una direzione di *anisotropia* nel piano XY . Invece la polarizzazione circolare ha spesso a che fare con una caratteristica geometrica un po' più complicata, che in genere si chiama *chiralità* e che, per esempio, è coinvolta in strutture che si distribuiscono spazialmente a forma di molla a spirale: in questo caso possono essere facilmente individuate delle proprietà geometriche che rimandano alla polarizzazione circolare sinistra o destra.

Le sorgenti di radiazione di interesse pratico per l'ottica (per esempio lampade e laser) hanno in genere caratteristiche di polarizzazione abbastanza ben definite. Anche se questa classificazione è molto grossolana e limitata in termini di casistica, in genere le lampade, in particolare quelle a filamento, producono una luce che è *non polarizzata*, o, per meglio dire, è polarizzata *random*: la direzione di polarizzazione cambia nel tempo in un modo che non può essere predeterminato facilmente, o, se preferite, la luce emessa contiene *tutte* le possibili direzioni di polarizzazione. Questo può essere visto come conseguenza del fatto che, specie quando a emettere è della materia che si trova in equilibrio termico ad alta temperatura (il filamento che, come vedrete in futuro, è spesso una buona approssimazione per un *corpo nero*), non è possibile individuare una direzione privilegiata nello spazio. Invece i laser producono quasi sempre (non sempre sempre) radiazione polarizzata *linearmente*, per motivi talvolta costruttivi e più spesso per ragioni collegate direttamente al loro funzionamento. Di norma, non esistono sorgenti di uso pratico che emettano direttamente luce polarizzata circolarmente, per ottenere la quale è necessario manipolare la radiazione della sorgente (per esempio polarizzata linearmente) con uno dei metodi che accenneremo tra breve.

La rilevanza della polarizzazione può essere compresa riflettendo su tanti aspetti, per esempio sui seguenti:

- il campo elettrico è un vettore, e dunque la sua direzione ha in genere un ruolo importante nel de-

terminare i fenomeni di interazione fra radiazione e materia;

- lo studio della polarizzazione emessa, oppure della risposta alla luce polarizzata, è un potentissimo sistema analitico nell'ambito della cosiddetta spettroscopia;
- a questo proposito, è storia interessante quella che riguarda la risposta alla luce polarizzata circolarmente di molte delle sostanze molecolari che costituiscono le basi della materia vivente: pare che esse siano prevalentemente "levogire", cioè che "sentano" in modo più efficace la radiazione polarizzata circolarmente in un dato verso, e a tutt'oggi le motivazioni di questa "asimmetria" sono ancora misteriose;
- esiste un'infinità di applicazioni per sistemi in grado di analizzare o manipolare la polarizzazione; accanto a dispositivi ottici molto raffinati, che magari incontrerete nella vostra futura carriera scientifica, sistemi del genere li avete probabilmente davanti agli occhi proprio ora, se state guardando queste note sullo schermo (piatto) di un qualsiasi dispositivo elettronico.

III. MANIPOLAZIONE E ANALISI DELLA POLARIZZAZIONE

Focalizziamo ora la nostra attenzione su due dispositivi per l'analisi (e misura), o manipolazione, della polarizzazione in uso nelle esperienze pratiche di laboratorio, precisamente il *polarizzatore lineare (polaroid)* e le *lamine ritardanti*.

Questi dispositivi sono costruiti con materiali la cui risposta ottica dipende dalla polarizzazione della radiazione che ci incide sopra. In termini generali, materiali dotati di simili caratteristiche si dicono *otticamente attivi* ed è possibile distinguere fra due categorie principali di attività ottica: dicroismo e birifrangenza.

Il *dicroismo* ha a che fare con l'assorbimento, e dunque la trasmissione, della radiazione attraverso il materiale. Nel caso più semplice (dicroismo lineare), che è quello che ci interessa, un materiale dicroico assorbe o fa passare l'onda che ci incide in maniera dipendente dalla direzione di polarizzazione dell'onda stessa, o, per meglio chiarire, dell'angolo compreso tra questa polarizzazione e un asse caratteristico (un *asse ottico*) del materiale dicroico.

La *birifrangenza*, invece, non tiene conto di fenomeni di assorbimento, ma agisce invece sulla *fase* dell'onda grazie al fatto che nei materiali birifrangenti la velocità di fase dipende dalla polarizzazione dell'onda stessa. Come potete facilmente rendervi conto ricordando la definizione di velocità di fase, questo significa che in un materiale birifrangente l'indice di rifrazione (reale) dipende dalla polarizzazione.

A. Dicroismo e polaroid

Il polarizzatore lineare a cui facciamo qui riferimento, chiamato gergalmente polaroid, dal nome commerciale di una famosa invenzione di circa un secolo fa, è un dispositivo che può essere modellato come una distribuzione spaziale di elementi in grado di assorbire la luce e allineati tra loro lungo una specifica direzione. Nell'invenzione originaria, e anche nella realtà odierna, almeno nella maggior parte dei casi, gli elementi assorbenti sono delle molecole di una qualche sostanza organica (spesso *cromofori*, cioè molecole di colorante). Queste molecole vengono disperse in una matrice polimerica, per esempio polivinilacol (PVA), che può essere stirata meccanicamente lungo una certa direzione. In seguito allo stiro, le molecole si allineano e si forma una sorta di sistema allineato di dipoli elettrici (questa descrizione è estremamente semplificata, ma può andare bene per i nostri scopi). Allora è evidente che un polaroid è un sistema che presenta un'anisotropia spaziale nell'assorbimento ottico.

Infatti, la radiazione polarizzata linearmente lungo l'asse di questi dipoli viene efficacemente assorbita, cioè viene trasmessa con forte attenuazione al di là dello strato di polaroid. Quella polarizzata linearmente in direzione ortogonale all'asse dei dipoli passa pressoché inalterata. Supponiamo di avere un'onda con campo elettrico di ampiezza E_0 e polarizzazione \hat{e} generica (sul piano XY) e immaginiamo che questa onda incida sul polaroid. Chiamando \hat{p} la direzione (sul piano XY) ortogonale a quella dell'asse dei dipoli del polaroid (questa direzione corrisponde a quella di uno dei due assi ottici del dispositivo, l'altro asse ottico essendo allineato con i dipoli), avremo che l'ampiezza del campo in uscita dal polaroid è data dalla proiezione $E_0 \hat{e} \cdot \hat{p} = E_0 \cos \theta$, con θ angolo compreso fra \hat{e} e \hat{p} . Poiché l'intensità di un'onda elettromagnetica è, come ricordato in precedenza, proporzionale al quadrato dell'ampiezza del campo, l'intensità in uscita, I , è legata a quella in ingresso, I_0 , dalla relazione

$$I = I_0 \cos^2 \theta, \quad (17)$$

che qualche volta si chiama *legge di Malus*: si vede facilmente come l'intensità trasmessa dal polaroid si annulli per $\theta = q\pi/2$, con q intero *dispari* e diverso da zero, e dunque ha una "periodicità" π .

Oltre all'intensità, il passaggio attraverso questo tipo di polarizzatore implica anche una manipolazione della direzione di polarizzazione, che all'uscita risulta allineata a \hat{p} , coerentemente con l'operazione di proiezione lungo tale direzione che abbiamo introdotto.

Un'interessante applicazione sperimentale di tutto ciò può essere realizzata usando due polaroid, 1 e 2, messi uno dietro l'altro, e una sorgente di luce polarizzata linearmente (un laser). Supponiamo inizialmente di usare il solo polaroid 2 (il più "lontano" dalla sorgente) e di ruotarlo in modo da minimizzare la trasmissione della luce. Questo si ottiene quando i dipoli del polaroid 2 sono allineati rispetto alla direzione di polarizzazione della sorgente laser, cioè quando $\hat{p}_2 \cdot \hat{e}_0 = 0$ (i simboli usati

dovrebbero risultare di immediata comprensione, in particolare \hat{e}_0 è la direzione di polarizzazione della sorgente), ovvero $\theta_2 = \pi/2$. Ora interponiamo il polaroid 1 tra sorgente e polaroid 2: all'uscita del polaroid 1 il campo avrà un'ampiezza $E_{0,1} = E_0 \hat{p}_1 \cdot \hat{e}_0 = E_0 \cos \theta_1$ e una direzione $\hat{e}_1 // \hat{p}_1$. Questa radiazione incide quindi sul polaroid 2, la cui rotazione è stata precedentemente aggiustata: all'uscita del polaroid 2 l'ampiezza del campo è

$$\begin{aligned} E_{0,2} &= E_{0,1} \hat{p}_2 \cdot \hat{e}_1 = E_0 \cos \theta_1 \hat{p}_2 \cdot \hat{p}_1 = E_0 \cos \theta_1 \cos(\theta_2 - \theta_1) \\ &= E_0 \cos \theta_1 \cos(\pi/2 - \theta_1) = (E_0/2) \sin(2\theta_1). \end{aligned} \quad (18)$$

Di conseguenza l'interposizione del polaroid 1 può far "riapparire" radiazione in uscita dal polaroid 2. Ricordando la relazione tra intensità e campi, si trova facilmente che all'uscita del polaroid 2 l'intensità vale

$$I_2 = (I_0/4) \sin^2(2\theta_1), \quad (20)$$

che si annulla per $\theta_1 = q\pi/2$, con q intero, e dunque ha una periodicità $\pi/2$.

La Fig. 4 mostra una rappresentazione schematica di due possibili esperimenti con uno (sopra) e due (sotto) polaroid, una sorgente (laser) polarizzata linearmente, e un rivelatore. Queste configurazioni sono quelle usate nell'esperienza pratica. Notate che in un polaroid ordinario "reale", specie se dotato di uno spessore sufficiente a maneggiarlo con facilità, la trasmissione non è mai completa neanche per polarizzazione ortogonale all'asse dei dipoli. Infatti un polaroid "reale" appare grigio, cioè semi-trasparente, all'occhio, poiché il materiale di cui è composto assorbe in ogni caso (per qualsiasi polarizzazione) una certa parte della radiazione incidente. La conseguenza pratica è che il termine I_0 che compare nelle Eqs. 17,20 è da intendersi come una frazione dell'intensità prodotta dalla sorgente. Inoltre, a causa dell'eventuale presenza di luce spuria raccolta dal fotorivelatore, nella pratica potrebbe essere necessario aggiungere alle Eqs. 17,20 un termine costante di offset.

B. Birifrangenza e lamine ritardanti

Come già anticipato, esistono materiali per i quali l'indice di rifrazione (reale) n dipende dalla direzione di polarizzazione della luce che vi incide. Se ricordate la definizione, valida per i casi che stiamo esaminando, $n = \sqrt{\epsilon_r}$, questo vuol dire che tali materiali hanno una costante dielettrica anisotropa (una descrizione accurata richiederebbe di sostituire lo scalare ϵ_r con un tensore). Avere materiali dielettrici trasparenti (altrimenti la luce non potrebbe passarvi dentro - e qui supponiamo di avere a che fare con materiali perfettamente trasparenti) anisotropi è generalmente molto semplice. Per esempio la maggior parte dei materiali trasparenti (o semi-trasparenti) che esistono in natura sono anisotropi, essendo anisotropa la loro struttura cristallina. Il quarzo, la calcite, la mica, materiali che possono essere estratti dalle miniere, presentano facilmente assi ottici ortogonali fra di loro (indicheremo qui le loro direzioni con $\hat{a}_{//}$ e \hat{a}_{\perp}) tali che

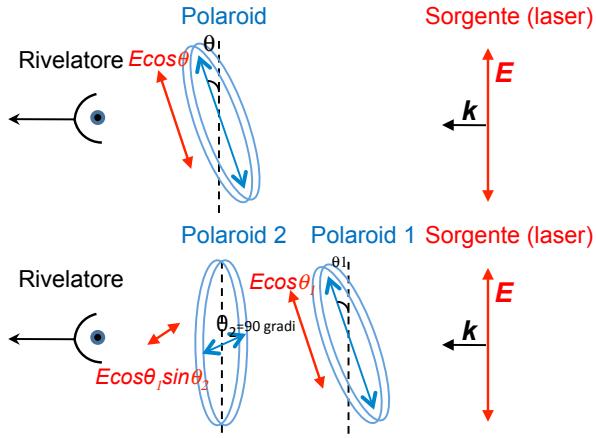


Figura 4. Illustrazione schematica di due possibili esperimenti con polaroid che fanno riferimento alle situazioni descritte nel testo e all'esperienza pratica.

l'indice di rifrazione è sensibilmente diverso per radiazione polarizzata lungo l'uno o l'altro di essi, comportando differenze $\Delta n = |n_{//} - n_{\perp}|$ anche dell'ordine di 0.1.

Una *lamina ritardante* (si può chiamare in tanti altri modi) è una lastra, generalmente sottile, di materiale birifrangente con caratteristiche dimensionali e di birifrangenza opportune per ottenere determinati scopi di manipolazione della polarizzazione. In particolare, le *lamine $\lambda/4$* servono per rendere circolare o ellittica una polarizzazione originariamente lineare (o viceversa) e le *lamine $\lambda/2$* servono per ruotare in modo controllato la direzione di una polarizzazione lineare.

1. Lamine $\lambda/4$

Supponiamo di avere un campo elettromagnetico di ampiezza E_0 e polarizzazione lineare lungo la bisettrice del piano XY , cioè con $\hat{e} = (\hat{x} + \hat{y})/\sqrt{2}$. La funzione d'onda corrispondente può essere scritta come

$$\vec{E} = \frac{E_0}{\sqrt{2}} \{(\exp[i(kz - \omega t)])\hat{x} + (\exp[i(kz - \omega t)])\hat{y}\}. \quad (21)$$

Supponiamo che questa onda incida su un materiale (trasparente) birifrangente con un certo Δn e un certo spessore d , e supponiamo anche che i due assi ottici del materiale siano paralleli rispettivamente agli assi X e Y . Scriviamo quindi la funzione d'onda che descrive la radiazione che emerge dal materiale stesso. Ricordando che $k = \omega/v = (\omega/c)n = k_0 n$, dovremo in questo caso introdurre due diversi numeri d'onda, $k_{//} = k_0 n_{//}$ e $k_{\perp} = k_0 n_{\perp}$, per tenere conto della birifrangenza. Notiamo anche che, per la scelta di orientamento che abbiamo fatto (serve solo per semplificare la matematica), $k_{//}$ è il numero d'onda per la componente di campo incidente polarizzato lungo uno dei due assi cartesiani, per esempio X , e k_{\perp} per la componente polarizzata lungo l'altro asse.

L'onda emergente ha la stessa forma della funzione di Eq. 21 e anche la stessa ampiezza, essendo il materiale trasparente, ma la coordinata z dovrà essere calcolata nel punto che corrisponde alla fine dello strato di materiale, cioè, per semplicità, in $z = d$:

$$\vec{E} = \frac{E_0}{\sqrt{2}} \{(\exp[i(k_{//}d - \omega t)])\hat{x} + (\exp[i(k_{\perp}d - \omega t)])\hat{y}\} \quad (22)$$

$$= \frac{E_0}{\sqrt{2}} \{(\exp[i(k_{//}d - \omega t)])\hat{x} + \exp[i(k_{\perp} - k_{//})d]\hat{y}\} \quad (23)$$

dove per l'ultima uguaglianza è stata fatta qualche semplice manipolazione algebrica.

In una lamina $\lambda/4$ si ha $(k_{\perp} - k_{//})d = k_0(n_{\perp} - n_{//})d = m\pi/2$, con m intero dispari. Notiamo che, visto che $k_0 = 2\pi/\lambda$, si ha anche $d(n_{\perp} - n_{//}) = m\lambda/4$, cioè la differenza di spessore ottico (prodotto fra indice di rifrazione e spessore fisico) è un multiplo dispari di $\lambda/4$, da cui la denominazione del dispositivo.

In queste condizioni si vede che la differenza di indice di rifrazione, e quindi di velocità di fase, delle componenti lungo le due polarizzazioni introduce uno sfasamento di $\pm\pi/2$ sulla componente Y . Quindi la radiazione, originariamente polarizzata lineare, diventa polarizzata circolare (destra o sinistra) a causa del passaggio attraverso la lamina. Inviando invece un'onda polarizzata circolarmente sulla lamina, in uscita si ottiene una polarizzazione lineare lungo la bisettrice del piano XY .

È facile infine verificare (ma qui non lo facciamo) che, se la polarizzazione incidente è diretta lungo una direzione che forma un angolo diverso da 45 gradi (la bisettrice del piano XY), allora la polarizzazione in uscita è *ellittica* con una certa orientazione e rapporto degli assi.

2. Lamine $\lambda/2$

In una lamina $\lambda/2$ si ha invece $(k_{\perp} - k_{//})d = k_0(n_{\perp} - n_{//})d = m\pi$, con m intero dispari, ovvero $d(n_{\perp} - n_{//}) = m\lambda/2$. Si verifica facilmente che in queste condizioni non si produce alcuno sfasamento tra le componenti, per cui l'onda in uscita dalla lamina resta polarizzata linearmente. Però la lamina agisce cambiando il segno di una componente rispetto all'altra: di conseguenza con una lamina a $\lambda/2$ la polarizzazione lineare diretta a 45 gradi rispetto agli assi viene cambiata di segno, con un effetto nullo sulla direzione.

Molto più interessante è verificare cosa succede se la radiazione incidente ha una polarizzazione lineare che forma un angolo ϕ generico (diverso da $\pi/4$) con un asse ottico del materiale birifrangente, per esempio con quello che abbiamo supposto parallelo all'asse X . In questo caso la funzione che descrive l'onda che emerge dalla lamina si scrive

$$\vec{E} = \frac{E_0}{\sqrt{2}} \{\cos \phi (\exp[i(k_{//}d - \omega t)])\hat{x} + \quad (24)$$

$$+ \sin \phi (\exp[i(k_{\perp}d - \omega t)])\hat{y}\} = \quad (25)$$

$$= \frac{E_0}{\sqrt{2}} (\exp[i(k_0 d - \omega t)]) (\cos \phi \hat{x} - \sin \phi \hat{y}), \quad (26)$$

che mostra come la polarizzazione in uscita sia ancora lineare, ma diretta lungo una direzione diversa da quella di ingresso. In particolare, tenendo conto del cambio di segno operato su una delle due componenti, si ha che si forma un angolo 2ϕ tra la direzione originaria e quella ottenuta in uscita dalla lamina, cioè la lamina $\lambda/2$ produce una rotazione di 2ϕ della direzione di polarizzazione lineare *senza attenuazione* dell'intensità (che è invece presente quando si usa un polaroid, o altro sistema dicroico).

La Fig. 5 mostra, in maniera molto schematica, due possibili e tipici impieghi delle lamine $\lambda/4$ (per convertire la polarizzazione lineare in polarizzazione circolare) e $\lambda/2$ (per ruotare la direzione di una polarizzazione lineare).

Infine è facile rendersi conto che una lamina ritardante “ideale” si comporta come ci si aspetta solo a una determinata lunghezza d’onda. Fortunatamente esistono dispositivi, generalmente di materiale plastico e quindi relativamente economici, che continuano a funzionare in modo ragionevole su un intervallo di lunghezze d’onda abbastanza esteso per permetterne l’uso pratico. Tuttavia, proprio a causa della possibilità di operare in un largo spettro di lunghezze d’onda, tali lamine, quando investite da una radiazione polarizzata linearmente, producono una polarizzazione generalmente ellittica.

Dal punto di vista pratico, la polarizzazione ellittica può essere qualitativamente individuata facendo seguire alla lamina ritardante un polaroid, e misurando l’intensità trasmessa dal polaroid. Ruotandolo, l’intensità non si annulla mai, a differenza di quanto avverrebbe con polarizzazione lineare. In particolare, se l’intensità misurata risulta indipendente dall’angolo di rotazione del polaroid, allora la polarizzazione emergente dalla lamina è perfettamente circolare.

IV. FRESNEL E BREWSTER

In ottica, un modo storicamente e concettualmente importante per manipolare la polarizzazione di una radiazione è quello che fa uso dell’incidenza al cosiddetto angolo di Brewster. In questo caso non si impiegano materiali dicroici o birifrangenti, ma si sfruttano le conseguenze delle regole che stabiliscono il comportamento di un’onda all’interfaccia tra due dielettrici con diverso indice di rifrazione, regole che qualche volta si chiamano *equazioni di Fresnel*.

Questo argomento fa normalmente parte di qualsiasi corso di “Fisica 2”, dove esso è anche contestualizzato a dovere. In questa nota ci limitiamo a ripercorrere rapidamente i concetti di base e i passaggi necessari per ricavare alcune delle equazioni di Fresnel, in particolare quelle che sono più strettamente connesse all’esistenza di un angolo di incidenza speciale, detto *angolo di Brewster*, per il quale la riflessione dall’interfaccia segue un andamento specifico per le varie componenti di polarizzazione.

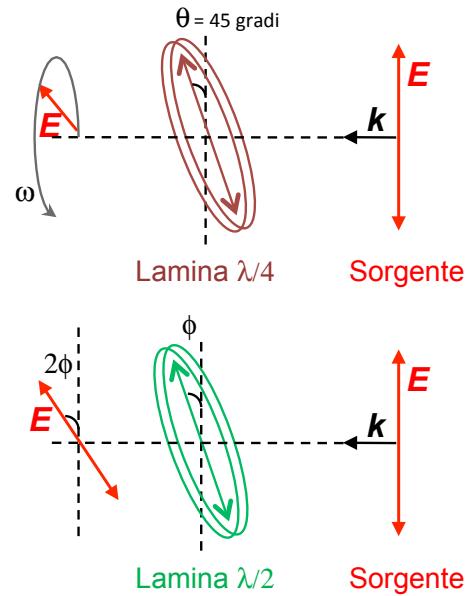


Figura 5. Illustrazione schematica per riassumere due situazioni di impiego pratico per lamine ritardanti.

A. Cenni alle equazioni di Fresnel e al procedimento per ricavarle

Supponiamo un’onda elettromagnetica che incide, propagandosi lungo la direzione \hat{k}_i , sull’interfaccia tra due materiali dielettrici con indice di rifrazione rispettivamente n_1 e n_2 (ovviamente diversi tra loro). La direzione di incidenza forma un angolo θ_i con la normale all’interfaccia nel punto di incidenza: la normale e \hat{k}_i appartengono a un piano che si chiama *piano di incidenza*. Supponendo che il materiale sia trasparente, o semi-trasparente, l’onda incidente darà luogo a:

- un’onda trasmessa, o rifratta, che si propaga nel mezzo con indice di rifrazione n_2 lungo la direzione \hat{k}_t e tale che, per la legge di Snell, tra angolo di trasmissione e di incidenza vale la relazione $\sin \theta_t / \sin \theta_i = n_1 / n_2$ (per salvaguardare la generalità delle nostre conclusioni supponiamo ovviamente che $\theta_i < \theta_{crit}$, con $\theta_{crit} = \arcsin(n_2/n_1)$ angolo critico, al di sopra del quale si ha “riflessione totale”);
- un’onda riflessa che si propaga nel mezzo con indice di rifrazione n_1 lungo la direzione \hat{k}_r e tale che l’angolo di riflessione è $\theta_r = \theta_i$.

Se ci pensate un attimo e osservate la Fig. 6, che riporta tutte le grandezze geometriche rilevanti per la spiegazione, potete facilmente rendervi conto che esistono due situazioni diverse a seconda che la polarizzazione dell’onda incidente appartenga al piano di incidenza [pannello (a) della figura] o sia ortogonale a questo [pannello (b)]. Le

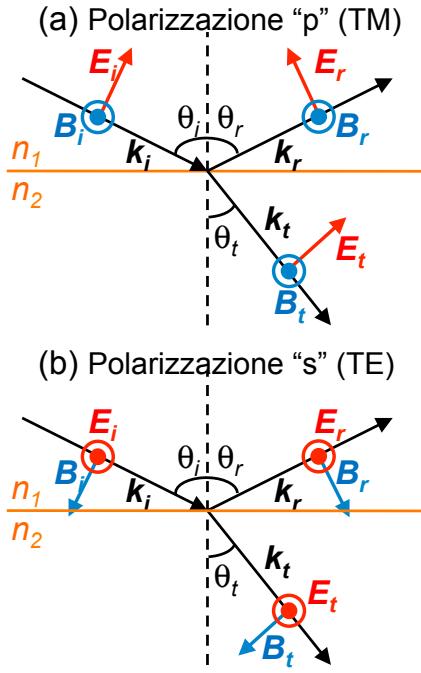


Figura 6. Geometria e simbologia utili per ricavare le equazioni di Fresnel nel caso di polarizzazione “p” (TM) e “s” (TE) [pannelli rispettivamente (a) e (b)].

due situazioni hanno un paio di denominazioni storiche rilevanti:

- se la polarizzazione è parallela al piano di incidenza si indica con la lettera “p”, altrimenti con la lettera “s”, dalle iniziali dei termini tedeschi che significano rispettivamente parallelo e ortogonale;
- se la polarizzazione è parallela al piano di incidenza, allora si indica l’onda come TM, a significare che il campo magnetico è trasverso al piano di incidenza, altrimenti si indica l’onda come TE, a significare che il campo elettrico è trasverso al piano di incidenza.

La Fig. 6 riporta anche, in forma grafica, diverse affermazioni che in gran parte possiamo ritenere ovvie, ma che richiederebbero dimostrazione (la potete trovare nei testi di elettromagnetismo e ottica). Per esempio: i tre vettori d’onda delle tre onde che stiamo considerando appartengono tutti al piano di incidenza, per ogni onda \hat{k} , \hat{e} , \hat{b} formano una terna ortogonale destrorsa e scegliamo di rappresentare i versi come rappresentato. Inoltre un’altra affermazione generale, anch’essa piuttosto ovvia, ma dimostrabile in modo rigoroso, è che nel passaggio da un mezzo all’altro la frequenza (o frequenza angolare) della radiazione non cambia, non essendo preso in considerazione nessun fenomeno che sia in grado di modificarla. Di conseguenza, è la lunghezza d’onda, ovvero il numero d’onda, che cambiano passando da un mezzo all’altro.

All’interfaccia devono valere le condizioni di continuità, o di raccordo, dei campi, che possono essere dedotte dall’integrazione delle equazioni di Maxwell. In particolare si conservano: $D_{\perp}, B_{\perp}, E_{//}, H_{//}$. Poiché siamo nel reame dell’ottica convenzionale e immaginiamo anche in questo caso che i materiali siano omogenei, isotropi, e “lineari”, la conservazione di D_{\perp} equivale a quella di $\epsilon_r E_{\perp}$, mentre quella di $H_{//}$ equivale alla conservazione di $B_{//}$ (e qui abbiamo usato la circostanza che $\mu = \mu_0$ dappertutto).

Le leggi di Fresnel si ottengono imponendo le condizioni di continuità all’interfaccia nel punto di incidenza. Dato che la frequenza della radiazione non cambia nel passaggio tra i mezzi e che si vuole, naturalmente, che le condizioni di continuità valgano *in ogni istante*, allora le conservazioni di cui sopra valgono anche tra le *ampiezze* dei vari campi coinvolti. Questo permette di semplificare la trattazione: infatti sappiamo che la relazione tra le ampiezze di campo elettrico e magnetico della stessa onda recita $E_0 = vB_0 = (c/n)B_0$, cosa che consente, quando necessario, di convertire le condizioni di continuità sulle componenti dei campi magnetici in condizioni di continuità sulle componenti dei corrispondenti campi elettrici.

B. Onda TM o polarizzazione “p”

Ci limitiamo a considerare la situazione descritta in Fig. 6(a), dato che questa è sufficiente per individuare l’esistenza dell’angolo di Brewster. Abbiamo a disposizione quattro condizioni di continuità sulle ampiezze, ma, essendo tre le incognite che ci interessano (le ampiezze E_{0i}, E_{0r}, E_{0t} delle onde incidente, riflessa, trasmessa), basta usarne due per determinare il legame tra le incognite e risolvere il problema che vogliamo trattare.

Dalla continuità di $H_{//}$, che è diventata continuità di $B_{//}$ e di qui continuità delle ampiezze di $nE_{//}$, abbiamo

$$n_1(E_{0i} + E_{0r}) = n_2E_{0t}, \quad (27)$$

che si può anche scrivere

$$E_{0i} + E_{0r} = E_{0t} \frac{n_2}{n_1}. \quad (28)$$

Dalla continuità di $E_{//}$ abbiamo invece, notando la geometria del problema e l’orientazione dei vettori che tiene conto della necessità di avere terne destrorse per le varie onde (da cui un non irrilevante segno meno),

$$E_{0i} \cos \theta_i - E_{0r} \cos \theta_r = E_{0t} \cos \theta_t, \quad (29)$$

che, notando che $\theta_i = \theta_r$, si può anche scrivere

$$E_{0i} - E_{0r} = E_{0t} \frac{\cos \theta_t}{\cos \theta_i}. \quad (30)$$

Combinando le Eqs. 28 e 30 si ottiene che il rapporto r_p tra le ampiezze dell’onda riflessa e di quella incidente, detto talvolta *riflettività in ampiezza*, è, per

polarizzazione “p”:

$$r_p = \frac{E_{0r}}{E_{0i}} = \frac{n_2 \cos \theta_i - n_1 \cos \theta_t}{n_2 \cos \theta_i + n_1 \cos \theta_t}. \quad (31)$$

Poiché in genere si è interessati a conoscere il rapporto tra le intensità, o potenze, delle onde, che, come già stabilito, dipendono dal quadrato delle ampiezze, conviene definire e determinare anche la *riflettanza* per polarizzazione “p”, R_p , tale che, per la conservazione del flusso di energia, $R_p + T_p = 1$ (con T_p ovviamente definita *trasmittanza*):

$$R_p = \left| \frac{E_{0r}}{E_{0i}} \right|^2 = \left| \frac{n_2 \cos \theta_i - n_1 \cos \theta_t}{n_2 \cos \theta_i + n_1 \cos \theta_t} \right|^2. \quad (32)$$

Per avere un’idea del valore numerico di R_p poniamoci nel caso, semplicissimo, di incidenza normale, cioè $\theta_i = \theta_t = 0$. È evidente che in questo caso le situazioni di polarizzazione p e s non sono distinguibili, per cui $R_p = R_s$, con ovvio significato dei simboli. Supponiamo allora di incidere normalmente su un’interfaccia aria-vetro ($n_2 = n_{vetro} \approx 1.5$, per il vetro ordinario, mentre $n_1 = n_{aria} \approx 1$), dove supponiamo di usare luce nel visibile, per esempio di colore rosso. Si ottiene $R_p = |(n_2 - n_1)/(n_2 + n_1)|^2 \approx 0.04$, cioè circa il 4% della potenza incidente viene riflessa. Questo è il motivo per cui, in certe situazioni di illuminazione, vediamo in maniera abbastanza nitida la nostra faccia riflessa dal vetro di una finestra.

Ricordando che, per la legge di Snell, si ha $n_2/n_1 = \sin \theta_i / \sin \theta_t$, e usando in modo opportuno le relazioni tra le varie funzioni trigonometriche, si ha che R_p si può scrivere nella seguente forma compatta (vi invito a verificare facendo tutti i passaggi del caso, che qui non riporto):

$$R_p = \left| \frac{\tan(\theta_t - \theta_i)}{\tan(\theta_t + \theta_i)} \right|^2. \quad (33)$$

Il numeratore della frazione non può mai azzerarsi, dato che, per Snell e assumendo $n_1 \neq n_2$, è sempre $\theta_i \neq \theta_t$. Però può esistere un angolo θ_B , detto *angolo di Brewster*, tale che per $\theta_i = \theta_B$ il denominatore tende a infinito, cioè $R_p \rightarrow 0$.

Per determinare univocamente l’angolo θ_B in funzione dei valori n_1 e n_2 occorre usare di nuovo la legge di Snell, che individua θ_t in funzione di θ_i . Il procedimento non è molto diretto. Si può partire notando che

$$\tan(\theta_i + \theta_t) = \frac{\tan \theta_i + \tan \theta_t}{1 - \tan \theta_i \tan \theta_t}, \quad (34)$$

per cui la condizione che stiamo cercando implica

$$1 = \tan \theta_i \tan \theta_t, \quad (35)$$

ovvero

$$\sin \theta_i \sin \theta_t = \cos \theta_i \cos \theta_t = \sqrt{(1 - \sin^2 \theta_i)(1 - \sin^2 \theta_t)}. \quad (36)$$

Facendo il quadrato di entrambi i membri e usando la legge di Snell si trova

$$\left(\frac{n_1}{n_2} \right)^2 = \frac{1}{\sin^2 \theta_i} - 1. \quad (37)$$

Questa equazione è soddisfatta, e dunque l’angolo θ_B è determinato, quando

$$\tan \theta_i = \tan \theta_B = \frac{n_2}{n_1}. \quad (38)$$

Dunque, supponendo come prima di esaminare l’interfaccia aria-vetro e usando luce visibile, si ha $\theta_B = \arctan(n_{vetro}/n_{aria}) \approx 56$ gradi.

1. Rilevanza e interpretazione fisica dell’angolo di Brewster

Riassumendo la tanta matematica che abbiamo svolto, possiamo affermare che abbiamo trovato che esiste un angolo di incidenza, determinato dal rapporto tra gli indici di rifrazione (reali) dei mezzi che formano l’interfaccia, tale che *se incidiamo con questo angolo la componente di polarizzazione parallela al piano di incidenza non viene riflessa*.

In termini pratici, se inviamo a questa interfaccia radiazione con polarizzazione random, come quella prodotta da una lampada a filamento, all’angolo di Brewster noteremo una sensibile diminuzione della intensità della radiazione riflessa, dato che in queste condizioni le componenti di polarizzazione p non saranno riflesse. Inoltre, e come diretta conseguenza, la radiazione riflessa contrerà solo componenti di polarizzazione s, per cui sarà polarizzata. Selezionando con un polaroid solo le componenti di polarizzazione p della sorgente o della radiazione riflessa, si osserverà che la riflessione si annulla pressoché completamente.

La possibilità di manipolare la radiazione tramite incidenza su un’interfaccia all’angolo di Brewster è sfruttata in molti dispositivi, per esempio nei laser a gas, dove essa serve a diminuire le perdite per riflessione da parte della cavità risonante (vedrete in seguito il significato di questa terminologia).

Un’applicazione curiosa è nell’impiego di lenti per occhiali da sole trattate con filtri polarizzatori. Infatti il nome polaroid richiama alla maggior parte delle persone il marchio di lenti da occhiali da sole, che, appunto, sono rivestite di uno strato di polaroid. Quando si va in barca a vela, si scia sulla neve, o si guida un’automobile, o un aeroplano, in un giorno di sole, si può rimanere abbagliati a causa dell’intenso riverbero, o riflesso, da superfici dielettriche piane collocate all’orizzonte (la strada, la neve, il mare, etc.), che hanno tutte indice di rifrazione tipicamente prossimo a quello del vetro. La direzione di osservazione di queste superfici, normalmente lontane dall’osservatore e collocate più o meno all’orizzonte, è più o meno all’angolo di Brewster, che dunque è anche l’angolo con cui la luce del sole, che può essere considerata polarizzata random, incide sull’interfaccia tra aria

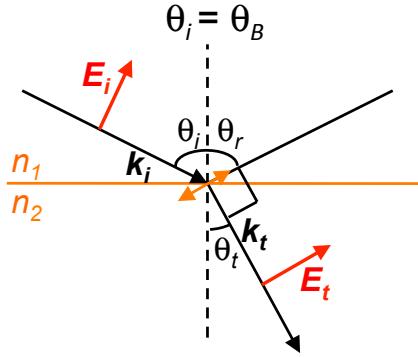


Figura 7. Rappresentazione schematica utile per dare un’interpretazione fisico/geometrica dell’esistenza dell’angolo di Brewster θ_B : per incidenza a questo angolo, nel solo caso di polarizzazione “p”, si ha che il dipolo all’interno del dielettrico con indice di rifrazione (reale) n_2 oscilla lungo la direzione indicata con la doppia freccia in figura. La riflessione tende ad annullarsi, dato che l’angolo di riflessione visto nel mezzo con indice di rifrazione n_1 corrisponde alla direzione dell’asse del dipolo oscillante, dove il pattern di radiazione è nullo.

a mezzo che riverbera. La riflessione contiene soprattutto le componenti di polarizzazione s: dunque se la lente contiene un filtro polarizzatore opportunamente orientato, e quindi in grado di assorbire, in maniera più o meno efficace, queste componenti, l’abbigliamento è attenuato.

Facciamo ora qualche considerazione geometrica utile per dare un’interpretazione fisica semplice da ricordare (molto più semplice di tutti i conti fatti sopra) per l’esistenza dell’angolo di Brewster. Abbiamo notato che per incidenza a questo angolo, cioè per $\theta_i = \theta_B$, si ha $\theta_i + \theta_t = \pi/2$. Come illustrato in Fig. 7, in queste condizioni si ha che le direzioni di propagazione dell’onda riflessa e dell’onda trasmessa, cioè \hat{k}_r e \hat{k}_t , formano un angolo retto (ricordate che $\theta_r = \theta_i$, per cui $\theta_i + \theta_t = \pi/2$ implica $\theta_r + \theta_t = \pi/2$, da cui l’affermazione fatta). L’onda riflessa e quella trasmessa possono essere considerate come generate dalla radiazione dei dipoli del materiale con indice di rifrazione (reale) n_1 eccitati dal campo elettrico dell’onda incidente. Poiché tali dipoli si trovano all’interno del materiale, essi vedono un’onda incidente la cui direzione di propagazione è stata rifratta, cioè è stata modificata rispetto a quella originaria in seguito al passaggio attraverso l’interfaccia. Dunque i dipoli vengono eccitati dal campo dell’onda che si propaga lungo \hat{k}_t . A causa dell’ortogonalità dell’onda elettromagnetica, tale campo ha una direzione di polarizzazione specifica, che in figura è rappresentata dalla doppia freccia. Per incidenza all’angolo di Brewster (e, ovviamente, solo nel caso di polarizzazione incidente p), la direzione in cui oscillano i dipoli coincide con quella di osservazione dell’onda riflessa. Ricordando che un dipolo elettrico oscillante non emette radiazione lungo la direzione del proprio asse, si ottiene una buona spiegazione del perché in queste condizioni l’onda riflessa si annulli.

$$r_s = \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t} = t_s - 1,$$

$$t_s = \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_i + n_2 \cos \theta_t} = r_s + 1,$$

$$r_p = \frac{n_2 \cos \theta_i - n_1 \cos \theta_t}{n_1 \cos \theta_t + n_2 \cos \theta_i},$$

$$t_p = \frac{2n_1 \cos \theta_i}{n_1 \cos \theta_t + n_2 \cos \theta_i}.$$

Figura 8. Quadretto riassuntivo per le riflettività r_s e r_p e trasmittività t_s e t_p in ampiezza per i due tipi di polarizzazione “s” e “p”.

C. Onda TE o polarizzazione “s” e quadro di sintesi

Un approccio simile a quello utilizzato in precedenza può essere applicato anche per determinare il rapporto tra le ampiezze dei campi, e quindi riflettività r_s e riflettanza R_s , nel caso di polarizzazione s, cioè con campo elettrico trasverso al piano di incidenza [Fig. 6(b)]. Non si riportano qui i passaggi e la tanta matematica necessaria, ma si cita solo il risultato finale, che è

$$R_s = \left| \frac{E_{0r}}{E_{0i}} \right|^2 = \left| \frac{n_1 \cos \theta_i - n_2 \cos \theta_t}{n_1 \cos \theta_i + n_2 \cos \theta_t} \right|^2 = \quad (39)$$

$$= \left| \frac{\sin(\theta_t - \theta_i)}{\sin(\theta_t + \theta_i)} \right|^2. \quad (40)$$

Si può verificare abbastanza facilmente come per questa polarizzazione non esista un angolo θ_i in grado di annullare la riflettività, dunque l’angolo di Brewster “non esiste” per polarizzazione s. Inoltre, dato che anche in questo caso il flusso di energia si conserva, cioè deve essere $1 = R_s + T_s$, si capisce come T_s sia una funzione piuttosto complicata e sicuramente non costante dell’angolo di incidenza, aspetto che può essere rilevante per capire l’osservazione sperimentale della trasmissione di luce polarizzata random attraverso un pacco di lastre dielettriche.

Per comodità, si riportano in Figs. 8 e 9 un quadretto riassuntivo per le riflettività r_s e r_p e trasmittività t_s e t_p in ampiezza per i due tipi di polarizzazione s e p, e il risultato di un calcolo numerico per le riflettanze R_s e R_p nel caso di singola interfaccia aria/vetro (materiale proveniente da wikipedia).

1. Effetti dell’angolo di Brewster in trasmissione

Gli effetti dell’esistenza dell’angolo di Brewster sono particolarmente evidenti quando si eseguono osservazioni in riflessione: il fascio riflesso di una luce non polarizzata che incide all’angolo $\theta_i = \theta_B$ non contiene la componente p di polarizzazione, dunque esso è polarizzato unicamente s.

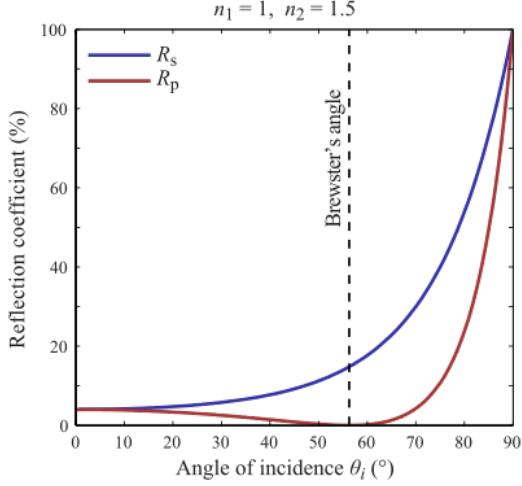


Figura 9. Risultato di un calcolo numerico per le riflettanze R_s e R_p nel caso di singola interfaccia aria/vetro.

Vediamo le caratteristiche di polarizzazione del fascio trasmesso (sempre supponendo che $\theta_i = \theta_B$). Per la conservazione dei flussi di energia si ha in generale per la trasmittanza

$$T_p = 1 - R_p = 1 - \left| \frac{\tan(\theta_t - \theta_i)}{\tan(\theta_t + \theta_i)} \right|^2 \quad (41)$$

$$T_s = 1 - R_s = 1 - \left| \frac{\sin(\theta_t - \theta_i)}{\sin(\theta_t + \theta_i)} \right|^2. \quad (42)$$

Per incidenza all'angolo di Brewster si ha $\theta_t + \theta_B = \pi/2$, per cui $T_p = 1$ e $T_s = 1 - |\sin(\theta_t - \theta_B)|^2 = \cos^2(\theta_t - \theta_B)$. Inoltre vale la legge di Snell, per cui $\theta_t = \arcsin(n_1 \sin \theta_B / n_2)$. Evidentemente la componente p, non riflessa all'interfaccia, viene completamente trasmessa, mentre quella s viene attenuata di un certo fattore.

Per avere un'idea di quanto valga questo fattore di attenuazione, stimiamo T_s nel caso di incidenza all'angolo di Brewster per un'interfaccia aria/vetro, dove $\theta_B \approx 56$ gradi e quindi $\theta_t \approx 34$ gradi. Di conseguenza $T_s \approx 0.85$, cioè circa l'85% della radiazione incidente polarizzata s viene trasmessa. Su una singola interfaccia è difficile rendersi conto a occhio dell'attenuazione provocata dal fenomeno. Però, se si pongono in serie numerose interfacce aria/vetro (o aria/PMMA, un materiale che ha un indice di rifrazione simile a quello del vetro), cioè se si usa un pacco di lastre dielettriche trasparenti, tutte ovviamente intervallate da sottili strati di aria, allora l'effetto complessivo potrà diventare osservabile. Infatti, detto m il numero di interfacce che il fascio incidente deve attraversare prima di essere osservato, la trasmissione totale per la componente polarizzata s è $T_{s,tot} = T_s^m$. Supponendo per esempio $m = 8$, si ha, nelle condizioni che stiamo trattando, $T_{s,tot} \approx 0.28$, mentre per $m = 12$ si ha $T_{s,tot} \approx 0.15$ e per $m = 16$ $T_{s,tot} \approx 0.08$. Se ricordiamo che, invece, la componente p attraversa le interfacce senza perdere intensità, possiamo concludere che all'uscita di un pacco di lastre dielettriche la luce non polarizzata e incidente all'angolo di Brewster sulla prima interfaccia risulta quasi (per oltre il 90% nel caso in cui si impieghino 16 interfacce) completamente polarizzata p.

Interferenza e diffrazione

francesco.fuso@unipi.it; <http://www.df.unipi.it/~fuso/dida>

(Dated: version 6 - FF, 4 maggio 2016)

Questa nota espone alcuni, *selezionati*, argomenti e concetti che hanno a che fare con l'interferenza e la diffrazione, con specifico riferimento al caso ottico (radiazione elettromagnetica nel visibile). Non c'è alcuna pretesa di completezza né di fronte alla ricchezza dei fenomeni coinvolti, né per quello che riguarda l'accuratezza della trattazione matematica. Chi è interessato può facilmente trovare nei testi di elettromagnetismo e ottica delle discussioni ben più complete e approfondite.

I. INTRODUZIONE

Interferenza e diffrazione sono due concetti intimamente connessi tra loro che hanno un'importanza fondamentale nell'ambito della meccanica ondulatoria. Le conseguenze di interferenza e diffrazione sono estremamente importanti soprattutto nell'ottica. In questo ambito, esse danno luogo a fenomeni ben noti e parecchio rilevanti, nei quali è spesso difficile distinguere tra ruolo specifico dell'interferenza e della diffrazione.

In termini molto qualitativi e descrittivi, l'interferenza è quel fenomeno che stabilisce una modulazione spaziale nell'intensità di un campo elettromagnetico che è dato dalla sovrapposizione di diverse onde. La diffrazione è invece più direttamente collegata alla modifica della distribuzione spaziale dell'intensità di un'onda che attraversa delle aperture di dimensioni trasversali limitate, ovvero "interagisce" con oggetti di piccole dimensioni.

II. INTERFERENZA

Facciamo riferimento a una situazione (molto) ideale: due sorgenti puntiformi, localizzate in due distinte posizioni sull'asse Z di un riferimento cartesiano, emettono onde piane monocromatiche *alla stessa frequenza* ω e con la stessa direzione e verso di propagazione (supponiamo coincidente con l'asse Z). Immaginiamo inoltre che, per semplicità, le onde si propaghino nel vuoto e che la loro polarizzazione sia lineare. Chiamiamo δ la distanza, misurata rispetto all'asse Z , delle due sorgenti. Possiamo scrivere le loro funzioni d'onda (supposte onde piane) come:

$$\vec{E}_1 = E_{01} \exp[i(kz - \omega t)]\hat{e}_1 \quad (1)$$

$$\vec{E}_2 = E_{02} \exp[i(k(z + \delta) - \omega t)]\hat{e}_2, \quad (2)$$

con ovvio e già noto significato dei simboli.

In ogni piano XY il campo sarà dato dalla *sovraposizione* (somma vettoriale) dei campi delle due onde, cioè $\vec{E} = \vec{E}_1 + \vec{E}_2$. Determiniamo l'*intensità* I di questo campo risultante. Ricordando che $I = <|\vec{S}|> = c\epsilon_0 |\vec{E}|^2/2$ (siamo nel vuoto), concentriamoci sul calcolo del modulo quadro del campo elettrico. Si vede facilmente che, supponendo le ampiezze delle onde reali,

$$|E|^2 = E_{01}^2 + E_{02}^2 + \hat{e}_1 \cdot \hat{e}_2 (E_1 E_2^* + E_1^* E_2) = \quad (3)$$

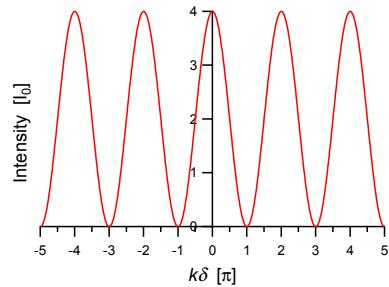


Figura 1. Grafico dell'Eq. 5 in funzione di $k\delta$.

$$= E_{01}^2 + E_{02}^2 + 2E_{01}E_{02}\hat{e}_1 \cdot \hat{e}_2 \cos(k\delta). \quad (4)$$

Al solo scopo di semplificare la matematica, poniamo anche $\hat{e}_1 = \hat{e}_2$ e $E_{01} = E_{02} = E_0$. In queste condizioni si ha $|E|^2 = 2E_0^2(1 + \cos(k\delta))$, ovvero

$$I = 2I_0(1 + \cos(k\delta)) = 4I_0 \cos^2(k\delta/2), \quad (5)$$

ancora con ovvio significato dei simboli e con l'uso, nell'ultima uguaglianza, di una semplice relazione trigonometrica. Notate che la procedura seguita per il calcolo del valore medio del modulo del vettore di Poynting è, di fatto, analoga a quella, usata altrove, in cui abbiamo posto $I = \text{Re}\{\vec{E} \times \vec{H}^*\}/2$, come si può facilmente verificare.

L'Eq. 5, rappresentata come funzione di $k\delta$ in Fig. 1, contiene il messaggio più importante del fenomeno dell'interferenza: *l'intensità di un'onda risultante dalla sovrapposizione di due (o più) onde è "modulata" in funzione della differenza di cammino (ottico)* [1] δ delle onde considerate nella sovrapposizione. In particolare l'intensità è massima quando $k\delta/2 = m\pi$, con m intero, ovvero, ricordando che $k = 2\pi/\lambda$, quando $\delta = m\lambda$, cioè la differenza di cammino ottico è pari a un multiplo intero di lunghezze d'onda. L'intensità è minima quando $k\delta = (2m + 1)\pi$, ovvero $\delta = (m + 1/2)\lambda$, cioè la differenza di cammino ottico è pari a un multiplo *dispari* di semilunghezze d'onda.

Nel caso considerato il *contrasto* (o visibilità) delle *frange di interferenza*, cioè il rapporto tra massimi e minimi di intensità, è massimo e vale uno. Tale contrasto può infatti essere definito come $(I_{max} - I_{min})/(I_{max} + I_{min})$, con ovvio significato dei simboli. È facile verificare (provateci) che, se le condizioni sulla polarizzazione e sull'in-

tensità delle onde vengono rilassate, l'interferenza continua a verificarsi (a meno che $\hat{e}_1 \perp \hat{e}_2$) con contrasto minore. Può essere un utile esercizio anche quello consistente nel verificare che si ha ancora interferenza, ma con spazialità e contrasto diverso, se i due vettori d'onda non sono collineari.

Sottolineiamo un paio di aspetti concettualmente rilevanti dell'interferenza. In primo luogo, nel fenomeno si ottiene che una grandezza scalare e stazionaria (cioè mediata nel tempo), ovvero l'intensità, viene a dipendere dall'argomento della funzione d'onda (qui compare δ). Inoltre, e questo è molto importante dal punto di vista delle applicazioni, la relazione che determina i massimi e i minimi di interferenza dipende dal rapporto λ/δ : pertanto sfruttando l'interferenza è possibile realizzare dei metodi che permettono di misurare delle distanze (δ) essendo nota λ o, viceversa, di misurare λ essendo nota δ .

Più in generale, l'intensità dell'onda ottenuta per sovrapposizione, che in questo caso abbiamo fatto dipendere da δ , è funzione della *differenza di fase*, o sfasamento, tra le onde che interferiscono. Supponiamo infatti di avere una diversa configurazione sperimentale, in cui poniamo $\delta = 0$, cioè le due sorgenti sono posizionate nello stesso punto. Supponiamo poi che negli argomenti delle funzioni d'onda di Eq. 1 siano presenti dei termini di fase costante, per esempio ϕ_1 e ϕ_2 , con $\Delta\phi = \phi_2 - \phi_1$. Ripetendo il procedimento, troveremmo che l'intensità è funzione proprio dello sfasamento $\Delta\phi$. Dovremmo infatti ottenere $I = 2I_0(1 + \cos(\Delta\phi))$. Ciò non deve stupire, poiché, in generale, la comparsa di termini di fase costante nell'argomento della funzione d'onda corrisponde a traslare l'origine del sistema di riferimento (delle posizioni e/o dei tempi).

A. Coerenza

Come già anticipato, la situazione esaminata è molto ideale. Abbiamo infatti scritto due funzioni d'onda relative alle onde prodotte dalle due *distinte* sorgenti, distanti δ l'una rispetto all'altra, immaginando di poter riferire agli stessi sistemi di riferimento temporale (e spaziale, nell'onda propagante posizione e tempo sono "mescolati" fra loro). Questo equivale a porre pari a zero lo *sfasamento* tra le due sorgenti. Specie quando si esaminano problemi di ottica, in cui le frequenze di oscillazione sono dell'ordine di $10^{14} - 10^{15}$ Hz, questa scelta non è affatto realistica. Infatti l'affermazione corrisponde a supporre che gli emettitori, ad esempio dipoli oscillanti, delle due sorgenti oscillino sempre *in fase* tra loro. Questi emettitori sono degli oggetti materiali e come tali risentono di processi (termici, collisionali, etc.) che hanno una natura statistica e che intervengono statisticamente per modificare la fase di un gruppo di oscillatori rispetto a un altro.

La situazione ideale immaginata corrisponde a dichiarare che le due sorgenti sono *coerenti* fra loro, ovvero,

appunto, che producono onde la cui relazione di fase, o sfasamento, rimane costante nel tempo. La caratteristica di *coerenza* è di estremo interesse in ottica, sia classica che quantistica, e i concetti che ad essa sono collegati possono avere diverse declinazioni, tutte intrecciate fra loro, a seconda del problema specifico che si vuole trattare. Dunque possono esistere diverse definizioni di coerenza (per esempio coerenza spettrale, temporale, spaziale, etc.) a seconda dello specifico problema che si sta affrontando. Naturalmente questa non è la sede giusta per discutere l'argomento in modo completo, e ci accontentiamo di vedere le conseguenze della coerenza in ambiti semplici, a partire dal problema (ideale) di interferenza che abbiamo citato.

In questo problema, una situazione molto più realistica è quella in cui le due sorgenti indipendenti sono rimpiazzate da due frazioni (in intensità) della stessa onda. Questo è quanto si verifica ad esempio nell'interferometro di Michelson (ma anche di Fizeau, di Fabry-Perot, di Bragg, etc.). In questo caso, le onde che si sovrappongono possono essere "automaticamente" in fase tra loro, essendo generate dalla stessa sorgente. D'altra parte, "frazionare" un'onda, cioè "dividerla" (fare uno *splitting*) in parti che corrispondono a onde che portano una frazione di intensità e hanno direzioni di propagazione diverse fra loro, è una procedura tecnicamente semplice. Infatti è sufficiente impiegare dei *beam splitters*, cioè degli specchi semi-riflettenti, per raggiungere lo scopo nella maggior parte delle configurazioni sperimentali di interferenza.

Tuttavia, come chiariremo con un esempio nella prossima sottosezione, operare in questo modo non garantisce di avere sovrapposizione di onde coerenti, a meno che le caratteristiche dell'unica sorgente che si impiega non siano adeguate rispetto all'esperimento che si vuole condurre. In altre parole, la caratteristica di coerenza che abbiamo qui attribuito a due distinte onde può essere applicata alla singola sorgente, che deve essere coerente affinché l'interferenza funzioni come richiesto.

1. Pacchetti d'onda e coerenza

Infatti c'è sicuramente un altro aspetto di debolezza concettuale nella descrizione usata in Sezione II. Sapete tutti che un'onda non può essere puramente monocromatica. Applicando il cosiddetto *principio di indeterminazione* (credo che tutti ne abbiate conosciuto almeno una formulazione, e forse lo avete anche già dimostrato matematicamente), per ottenere la pura monocromaticità occorre supporre che la sorgente sia stata accesa a un tempo infinitamente precedente a quello di osservazione e che venga spenta a un tempo infinitamente successivo.

Più realisticamente dovremmo considerare un'onda come composta da diverse *componenti spettrali*, cioè come data dalla sovrapposizione di onde dotate di frequenze diverse all'interno di un certo intervallo $\Delta\nu$, che possiamo chiamare grossolanamente *larghezza di riga*. Sappiamo già come comportarci nel caso di fenomeni ondulatori

periodici, dove è possibile scrivere l'onda risultante sotto forma di *serie* di Fourier di diverse componenti, ognuna di frequenza multiplo di una frequenza di base. Qui, a causa del fatto che le varie componenti hanno frequenza che varia in modo *continuo* nell'intervallo $\Delta\nu$, la serie è sostituita da un *integrale*.

Siete fortemente invitati a prendere Python e sommare fra di loro tante (virtualmente infinite) onde armoniche di frequenze leggermente (virtualmente infinitesimamente) diverse fra loro: vedrete che il risultato della somma avrà un'estensione temporale finita $\Delta t \sim 1/\Delta\nu$, cioè sarà sensibilmente diverso da zero in questo intervallo, come rappresentato in Fig.2: questo Δt in certi contesti ha il ruolo di misura dell'*intervallo temporale di coerenza* della sorgente. Tenendo conto della propagazione, che avviene alla velocità di fase dell'onda (velocità della luce, nel caso che stiamo esaminando in cui la propagazione è nel vuoto), a questo Δt corrisponde un'estensione spaziale finita, $\Delta L \sim c\Delta t = c/\Delta\nu$. A questo ΔL in certi contesti viene attribuito il nome di *lunghezza di coerenza* della sorgente. Dunque un $\Delta\nu$ (relativamente) piccolo corrisponde a (relativamente) grandi valori di Δt e ΔL , cioè caratterizza un'onda (relativamente) molto coerente.

La sovrapposizione di tante onde armoniche produce dei *pacchetti d'onda* (un pacchetto d'onda è quello rappresentato in Fig. 2), oggetti matematici che vi saranno di grandissima utilità andando avanti con gli studi. In un esperimento in cui si vuole creare interferenza dividendo in due (o più) frazioni di intensità un'onda, come in un interferometro, è chiaro che l'interferenza può verificarsi solo dalla sovrapposizione degli "stessi" pacchetti d'onda. Pensate infatti all'interferometro di Michelson, e immaginate che la differenza di cammino ottico tra le due frazioni di onda che si vanno a sovrapporre sia maggiore dell'estensione spaziale dei pacchetti prodotti dalla sorgente: i due pacchetti d'onda che percorrono i due distinti cammini ottici nell'interferometro non si potranno sovrapporre, o si sovrapporranno solo parzialmente, sullo schermo o sul rivelatore impiegato per misurare l'intensità. Di conseguenza non si avrà interferenza, o il contrasto delle frange sarà ridotto e la visibilità del fenomeno sarà scarsa.

Come già sottolineato, sorgenti molto monocromatiche, cioè con un $\Delta\nu$ piccolo, producono pacchetti d'onda di grande estensione spaziale e temporale. Talvolta (e senza entrare troppo nei dettagli delle definizioni) si dice che tali sorgenti sono *spettivamente, spazialmente e temporalmente coerenti*. Un ottimo esempio è rappresentato dai laser, specie di alcune tipologie, all'interno dei quali i singoli emettitori (ad esempio, dipoli elettrici) sono in qualche modo costretti ad oscillare tutti in fase tra loro fornendo una radiazione coerente. In alcuni laser si ottengono piuttosto facilmente larghezze di riga $\Delta\nu < 100$ Hz (a questo corrisponde un rapporto $\Delta\nu/\nu \sim 10^{-12}$, che sancisce in modo chiarissimo la "superiorità" dei laser rispetto a qualsiasi altra sorgente di radiazione in termini di "accuratezza"), che danno luogo nel vuoto a pacchetti d'onda estesi qualche migliaio di chilometri. I laser a

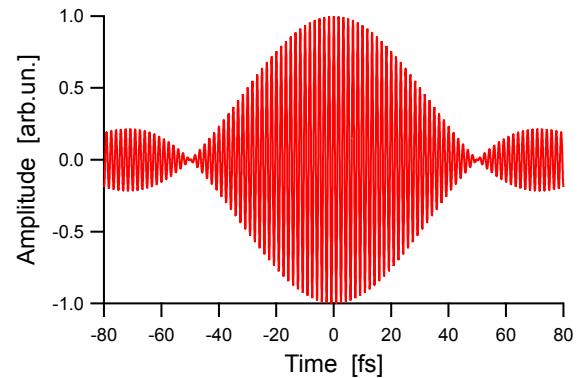


Figura 2. Risultato della somma di mille funzioni armoniche del tempo di frequenza omogeneamente distribuita in un intervallo $\Delta\nu = 2 \times 10^{13}$ Hz centrato attorno alla frequenza $\nu = 6 \times 10^{14}$ Hz (onda molto poco monocromatica!). Si osserva la formazione di un "pacchetto d'onda" nel dominio del tempo, la cui estensione temporale è $\Delta t \approx 2/\Delta\nu = 100$ fs.

diodo che usate in laboratorio non sono propriamente un ottimo esempio di sorgente coerente: più che la larghezza di riga (che è in genere dell'ordine dei MHz), conta la scarsa stabilità di operazione a tempi medi/lunghi e la possibilità, tipica proprio dei laser a diodo, di "saltare" in modo discreto da una frequenza a un'altra. Di fatto i pacchetti d'onda generati difficilmente hanno estensioni spaziali maggiori di qualche decina di cm. Infine, un veramente pessimo esempio di sorgente coerente è rappresentato da una lampada a filamento, dove i singoli emettitori si comportano ognuno per conto suo (l'emissione ha un'origine termica, inerentemente stocastica) e la lunghezza di coerenza è molto molto piccola.

III. HUYGENS E DIFFRAZIONE

Prima di procedere con la descrizione di alcuni fenomeni di diffrazione e di discutere come essi siano legati all'interferenza, è necessario richiamare alcuni principi e teoremi che sono parte dell'ottica ondulatoria e che in questa nota saranno solo brevemente citati per sommi capi. In particolare ci serve quello che in genere viene chiamato *principio di Huygens*, che a sua volta deriva da un teorema detto di Kirchoff (o di Kirchoff-Fresnel, o forse anche di qualcun altro). Ci serve sapere che *ogni piccola regione di un fronte d'onda si comporta come una sorgente di onde secondarie*; queste onde secondarie sono tutte in fase tra loro, hanno la forma di onde sferiche e sono prevalentemente emesse nello stesso verso e direzione dell'onda primaria, quella di cui stiamo considerando il fronte d'onda.

Teoremi e principi di questo tipo sono stati importantissimi nella storia dell'ottica, in particolare nell'800, per dimostrare matematicamente la possibilità di propagazione di un'onda elettromagnetica nel vuoto, costituendo un valido elemento di opposizione ai teorici dell'etere. Al

giorno d'oggi del concetto di etere non c'è più bisogno, e molte delle affermazioni contenute in questi teoremi e principi suonano piuttosto ovvie. Come vedremo tra breve, però, c'è qualcosa che rende molto utile servirsi di questi principi nell'ambito di quello che vogliamo analizzare.

A. Interferenza da doppia fenditura (Young)

Ci serviamo del famoso esperimento della doppia fenditura (esperimento di Young) per introdurre il passaggio da interruzione a diffrazione. La fama di questo esperimento si deve soprattutto al fatto che esso è in genere considerato come un ottimo banco di prova per verificare il dualismo onda/particella (e tanti altri bellissimi argomenti), e lo incontrerete di sicuro in futuro proprio in questa veste. Qui siamo completamente ondulatori e quindi ci limitiamo a interpretare l'esperimento come interferenza fra due onde. Notate che, al termine di queste note, torneremo sullo stesso esperimento per delle precisazioni molto rilevanti.

Nell'esperimento di Young un'onda, che supponiamo piana e monocromatica (abbiamo già accennato a come si fa a trattare situazioni un po' più realistiche, ma non vogliamo complicarci la vita), incide ortogonalmente su una lamina opaca su cui sono praticate due (piccole) aperture lineari, chiamate anche *fenditure*. Le dimensioni trasversali di queste fenditure sono molto importanti per la descrizione completa del fenomeno, come vedremo al termine di questa nota. Per il momento supponiamo che esse siano "molto piccole" e che la *separazione* spaziale tra di esse sia d . Di fronte alla lamina, parallelamente e a grande distanza ($D \gg d$) da questa, si trova uno schermo su cui si osservano delle frange di interferenza caratterizzate da una certa intensità I dipendente dalla posizione angolare $\sin \theta$. La Fig. 3(a) mostra uno schema dell'esperimento.

Secondo il principio di Huygens le due aperture, che intercettano un fronte dell'onda "primaria", si comportano da sorgenti di onde secondarie le quali risultano in fase tra loro; queste due sorgenti si trovano inoltre in posizioni diverse nello spazio. Le onde prodotte dalle aperture sono sferiche, quindi emesse in tutte le direzioni (in "avanti", cioè verso lo schermo), però, visto che siamo interessati a vedere cosa succede sullo schermo, che si trova a grande distanza dalle sorgenti stesse, esse potranno essere bene approssimate da onde piane. Dunque in ogni posizione dello schermo c'è sovrapposizione delle due onde (coerenti) provenienti dalle aperture e di conseguenza si ha un fenomeno di interferenza.

Sceglieremo un punto generico sullo schermo: la sua coordinata sarà x (l'asse X è verticale in figura e centrato sull'"asse geometrico del sistema", che in questo caso passa per il punto di mezzo tra le due aperture). In alternativa alla coordinata x , potremo identificare il punto attraverso l'angolo θ , formato tra il raggio vettore e l'asse del sistema, la cui direzione coincide con quella di propa-

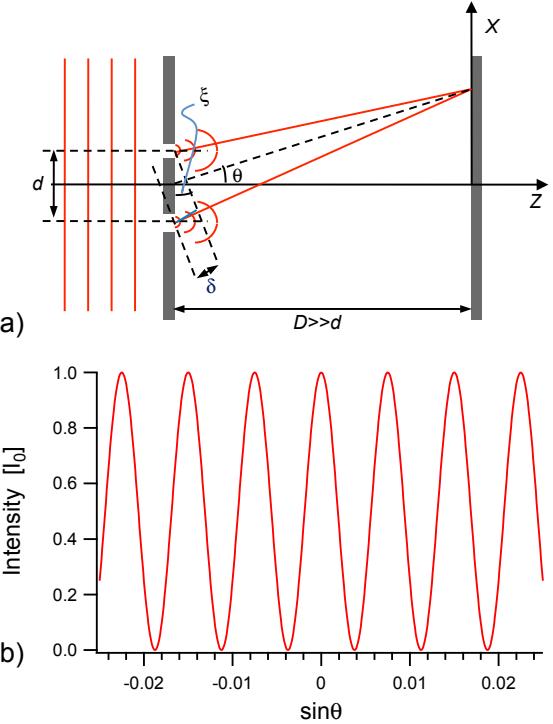


Figura 3. Schema dell'esperimento della doppia fenditura (a) e calcolo dell'intensità sullo schermo in funzione di $\sin \theta$ (b). Nel calcolo si è assunta una spaziatura $d = 100\mu\text{m}$ e una lunghezza d'onda $\lambda = 500 \text{ nm}$. Per chiarezza, si è considerato un piccolo intervallo di variazione di $\sin \theta$ e sono state applicate tutte le approssimazioni citate nel testo.

gazione dell'onda primaria. Questa descrizione a parole è complicata, per cui conviene riferirsi alla figura.

Sappiamo che l'interferenza costruttiva e distruttiva, ovvero la presenza di massimi e minimi di intensità, dipende dalla differenza di cammino ottico δ tra le due onde, ovvero dal loro sfasamento. Una semplicissima costruzione geometrica ci permette di individuare il segmento δ (vedi figura). Una altrettanto semplice considerazione ci permette di affermare che, per $D \gg d$, $\delta = d \sin \xi$, con ξ indicato in figura. Inoltre $\xi \approx \theta$, per cui $\delta \approx d \sin \theta$ (tutte le considerazioni geometriche svolte richiedono di immaginare una figura disegnata davvero in scala, cioè con D molto più grande di d).

A causa dell'interferenza, l'intensità $I(\theta)$ sullo schermo segue l'andamento descritto dall'Eq. 5, con $\delta \approx d \sin \theta$, cioè:

$$I(\theta) = I_0 \cos^2(kd \sin \theta / 2), \quad (6)$$

dove I_0 è qui il massimo dell'intensità della radiazione che incide sullo schermo e abbiamo "accettato" l'approssimazione sostituendo \approx con $=$. La Fig. 3(b) mostra l'andamento dell'intensità in funzione di $\sin \theta$: si osservano dei massimi e minimi di interferenza regolari. I massimi si trovano nelle posizioni angolari $(\sin \theta)_{max}$ tali che $kd(\sin \theta)_{max} = 2m\pi$, con m intero, cioè $(\sin \theta)_{max} =$

$2m\pi/(kd) = m\lambda/d$; i minimi si trovano nelle posizioni angolari $(\sin \theta)_{min}$ tali che $kd(\sin \theta)_{min} = 2(m+1)\pi$, con m intero, cioè $(\sin \theta)_{min} = (2m+1)\pi/(kd) = (m+1/2)\lambda/d$.

Osservate che, nelle tipiche condizioni sperimentali in cui $D >> d$, la trigonometria permette di fare le seguenti ulteriori approssimazioni: $\sin \theta \approx \theta \approx \tan \theta = x/D$. In altre parole, considerando piccole variazioni angolari e spaziali, che sono quelle di interesse, si ha che la *separazione spaziale* dei massimi o dei minimi sullo schermo (la separazione è la stessa per i massimi e i minimi) è $(\Delta x)_{max \text{ o } min} = \lambda D/d$: ritroviamo ancora una volta il legame tra posizione delle frange di interferenza con lunghezza d'onda e distanza (in questo caso separazione d fra le fenditure) che caratterizza in termini “generali” l'interferenza. Inoltre notiamo come sull'asse geometrico del sistema, in una posizione che è “schermata” geometricamente rispetto all'onda incidente, troviamo un bel massimo di interferenza, circostanza che potrebbe sembrare poco probabile se non si considerasse nei dettagli il fenomeno descritto.

B. Interferenza da reticolo ottico

Un'estensione particolarmente rilevante, soprattutto a causa delle notevoli applicazioni pratiche, è quella che prevede di rimpiazzare il sistema delle due fenditure con un sistema di tante, N , fenditure, tutte spaziate fra loro in modo regolare di una distanza d . Un sistema di questo tipo si chiama spesso *reticolo ottico* o, meglio, *reticolo di diffrazione* (concetti e denominazioni di interferenza e diffrazione cominciano a sovrapporsi tra loro anche nella nomenclatura), e la modalità di impiego a cui facciamo qui riferimento è detta *in trasmissione*, per distinguere da quella *in riflessione*, in cui i reticolli trovano applicazione ancora più ampia.

Ogni coppia di fenditure si comporta come nell'esempio precedente, cioè è sorgente di onde secondarie in fase tra loro. È evidente che per determinare la funzione dell'intensità $I(\theta)$ (θ è definito in analogia con prima) bisogna considerare l'interferenza tra tante onde. La matematica è complicata, e implica la convergenza di una serie non banale, come potete trovare in alcuni testi di ottica. Qui ci accontentiamo di riportare il risultato finale e i commenti che ci si possono fare sopra. Si ottiene:

$$I(\theta) = I_0 \frac{\sin^2(N\gamma)}{\sin^2 \gamma} \quad (7)$$

$$\gamma = \pi \frac{d}{\lambda} \sin \theta, \quad (8)$$

dove tutti i simboli sono ovvi o già definiti.

Vediamo le proprietà della funzione che abbiamo scritto. Cominciamo con il notare che, quando il numeratore e il denominatore tendono entrambi a zero, come si verifica per $\gamma \rightarrow m\pi$, con m intero, la funzione ha dei massimi (assoluti) che tendono al valore N^2 . Tenendo conto dell'espressione di γ , si vede come la condizione implichi $\sin(\theta_{max'}) = m\lambda/d$. Quindi si trova un massimo

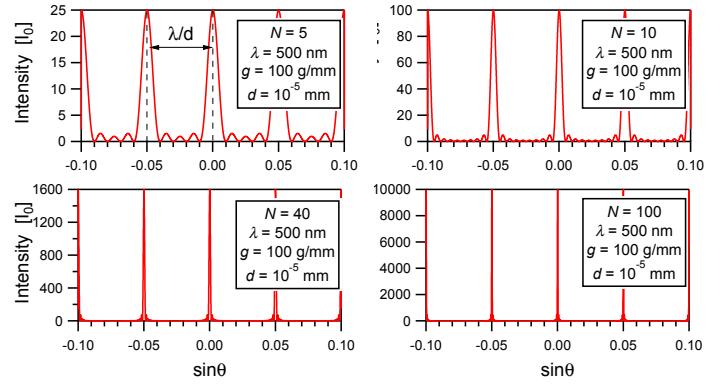


Figura 4. Calcolo della trasmissione attraverso un reticolo di diffrazione eseguito secondo l'Eq. 7. Nel calcolo si è supposto $\lambda = 500 \text{ nm}$ e $d = 1/g = 100 \mu\text{m}$. I diversi grafici si riferiscono a diversi valori di N , come in legenda.

“centrale”, per $m = 0$, e tanti altri massimi spaziati di λ/d . L'apice impiegato nelle espressioni, quello posto su max , è dovuto al fatto che esiste anche un'altra “tipologia” di massimi, che si hanno quando il numeratore ha un massimo, cioè $N\gamma = (m+1/2)\pi/2$ e, contemporaneamente, il denominatore è diverso da zero. Tenendo conto dell'espressione di γ , questi massimi, che hanno carattere relativo, si trovano per $(\sin \theta)_{max''} = [(m+1/2)/N]\lambda/d$. Tra un massimo relativo e il successivo ci sono ovviamente dei minimi, che si ottengono quando il numeratore va a zero e, contemporaneamente, il denominatore è diverso da zero. Questo si ottiene per $(\sin \theta)_{min} = (m/N)\lambda/d$. Per capire meglio l'andamento della funzione, conviene riferirsi a grafici di $I(\theta)$ calcolati numericamente, come quelli rappresentati in Fig. 4, dove il calcolo è svolto per diversi valori di d .

Come già affermato, la distanza, o spaziatura, fra due massimi (consecutivi) della prima tipologia, i massimi assoluti, vale $(\Delta \sin \theta)_{max'} = \lambda/d$ e non dipende dal numero N di aperture. Poi ci sono dei minimi, pari a zero, la cui spaziatura è $(\Delta \sin \theta)_{min} = \lambda/(Nd)$. Facendo un po' di matematica, si ottiene che tra un massimo assoluto e l'altro ci sono $(N-1)$ minimi. Infine è evidente che tra un minimo e l'altro si situano i massimi relativi che corrispondono alla tipologia indicata sopra con il doppio apice. Tra due massimi principali (assoluti) ci sono $(N-2)$ massimi relativi.

È molto interessante osservare come si modifica il sistema di frange di interferenza all'aumentare del numero N di aperture, o fenditure. Si vede in Fig. 4 che non solo il numero di minimi aumenta, ma anche che i massimi relativi diventano sempre più deboli mentre quelli principali si rinforzano (l'intensità di picco scala con N^2) e diventano sempre più snelli.

Nella pratica si hanno spesso a disposizione dei reticolli ottici di grandi dimensioni. Dunque aumentare il numero N di aperture interessate dalla radiazione significa di fatto aumentare le dimensioni del fascio di luce che in-

cide sul reticolo. Ne risulta la soppressione dei massimi relativi, l'aumento e lo "strizzamento" di quelli principali. Se immaginate di usare un reticolo di diffrazione per disperdere la radiazione (in modo molto più efficiente che non con un prisma) e misurare la lunghezza d'onda della radiazione, cosa ben possibile supponendo che d sia noto, allora è evidente che allargare la regione di reticolo illuminata porta notevoli vantaggi nell'accuratezza e sensibilità della misura.

Facciamo ancora qualche altra osservazione di tipo pratico. Spesso anche usando i reticolati di diffrazione si misura in realtà non lo spostamento angolare θ , ma quello lineare x , esattamente come nel caso dell'interferenza da doppia fenditura. A questo scopo si può spesso usare l'approssimazione $\sin \theta \approx x/D$, con D , al solito, distanza del reticolo dallo schermo di osservazione. Dunque la posizione dei massimi principali rispetto all'asse geometrico del sistema è $(x)_{max'} = mD\lambda/d$. All'intero m si dà spesso il nome di *ordine di diffrazione* e, per i motivi che chiariremo al termine di questa nota, in genere si preferisce, o talvolta si è costretti, a usare il primo ordine di diffrazione, $m = \pm 1$.

Sempre dal punto di vista pratico, notate che normalmente i reticolati vengono caratterizzati non con la spaziatura d tra le fenditure, ma con il suo reciproco, $g = 1/d$, che è la densità lineare di fenditure. In genere g si dà in *numero di righe* per mm. Per motivi legati alla costruzione e alla maggiore facilità con cui si riesce a illuminare una vasta porzione del reticolo, molto spesso i reticolati si usano *in riflessione*. Reticoli in riflessione da 1800 o anche 3600 righe/mm (o grooves/mm) sono piuttosto comuni in ottica, dove costituiscono l'elemento principale di strumenti diffusissimi per la spettroscopia (cioè per misurare le componenti spettrali dell'emissione) che si chiamano spettrometri o, spesso, *monocromatori*. Alcuni monocromatori, che fanno uso di reticolati molto grandi e di una grande "distanza focale" (l'equivalente della distanza D tra reticolo e schermo), permettono di risolvere emissioni con lunghezze d'onda separate da molto meno di 1 Å, come necessario per esempio nella cosiddetta spettroscopia Raman.

Ultimissima annotazione. Ci sono molti oggetti che si comportano in modo simile ai reticolati. Senza citare le ali della farfalla (che danno luogo all'iridescenza) o alcune vernici di automobili che ne simulano il funzionamento, e lasciando da parte anche le chiazze di olio sull'acqua, o in generale i fenomeni che si verificano quando ci sono degli strati sottili di materiali dielettrici sovrapposti, per i quali è più corretto fare riferimento all'interferenza (multipla) "alla Bragg", che studierete in altri contesti, possiamo ricordare CD e DVD. Essi sono (o erano) realizzati premarcando un substrato plastico con delle piste tangenziali, il cui pitch (distanza in direzione radiale fra una pista e l'altra) vale rispettivamente 1.6 e 0.74 μm. In certe condizioni queste piste possono comportarsi come le righe di un reticolo di diffrazione in riflessione (il substrato è opaco, e quindi in trasmissione non funziona), con una separazione pari al pitch. Infatti tutti sapete che,

osservando la riflessione della luce con spettro continuo (quella di una lampadina, per intenderci) su un CD o un DVD, si vedono i colori dell'iride separati spazialmente, ovvero dispersi, tra loro.

IV. DIFFRAZIONE DA SINGOLA FENDITURA (LINEARE)

Finalmente arriviamo a esaminare un caso in cui è davvero più opportuno parlare di diffrazione che non di interferenza, anche se l'interferenza è sempre un concetto da tenere ben presente per l'interpretazione del fenomeno.

Immaginiamo allora di avere una *singola fenditura* (di forma lineare) con una dimensione trasversale (apertura) a incisa su una lamina opaca. Suddividiamo questa apertura in tanti, virtualmente infiniti, elementini, di dimensioni trasversali virtualmente infinitesime. Applichiamo quindi il principio di Huygens a questi piccoli elementini: essi diventeranno sorgente di onde secondarie tutte in fase tra di loro. Se preferite, in una visione un po' più "fisica", si crea in questo modo un array di emettitori individuali, per esempio dipoli oscillanti, tutti in fase fra loro.

Quello che qui stiamo esaminando somiglia molto a un reticolo in cui abbiamo fatto tendere il numero N a infinito e la spaziatura d a zero, facendo in modo che il prodotto Nd tendesse ad a . Questo approccio, con molte cautele dovute alle complicazioni matematiche coinvolte, può essere utile per ricordare alcune caratteristiche della funzione $I(\theta)$ in questo caso. Con un po' di passaggi, non tutti banali (al solito, potete cercarli nei testi di ottica o elettromagnetismo), si ottiene la cosiddetta *funzione di diffrazione*:

$$I(\theta) = I_0 \frac{\sin^2(\alpha)}{\alpha^2} \quad (9)$$

$$\alpha = \pi \frac{a}{\lambda} \sin \theta , \quad (10)$$

dove tutti i simboli sono ovvi o già definiti.

Facendo il limite per $\alpha \rightarrow 0^\pm$ si vede che per questo valore la funzione ha un massimo assoluto, che corrisponde a $(\sin \theta)_{max'} = 0$. Si nota poi che la funzione ha dei minimi quando il numeratore si annulla (escluso, ovviamente, $\alpha \rightarrow 0^\pm$ che abbiamo appena riconosciuto come massimo), cioè in corrispondenza di $(\alpha)_{min} = m\pi$, con m intero, ovvero $(\sin \theta)_{min} = m\lambda/a$. Tra i minimi si trovano dei massimi relativi, che si hanno per $(\alpha)_{max''} = (m+1/2)\pi$, con m intero, ovvero $(\sin \theta)_{max''} = (m+1/2)\lambda/a$. Si capisce facilmente come la situazione sia ben diversa da quella del reticolo per quanto riguarda l'altezza dei massimi. Infatti il massimo assoluto è uno solo e tutti gli altri hanno un'altezza che va diminuendo all'aumentare di $|m|$. Tutto questo è ben riassunto nella Fig. 5, in cui sono stati riportati i risultati del calcolo per diversi valori di a .

La figura mostra in maniera molto chiara l'effetto eclatante della diffrazione: a parte la presenza dei massimi

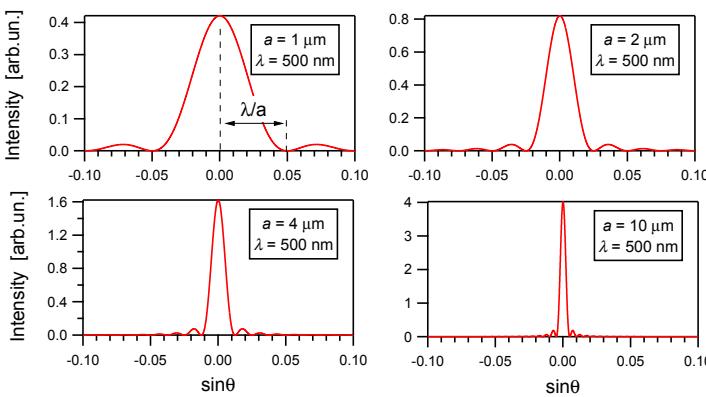


Figura 5. Calcolo della trasmissione attraverso una fenditura (lineare) di apertura a come in legenda, usando l'Eq. 9. In tutti i casi si è ipotizzata una radiazione di lunghezza d'onda $\lambda = 500 \text{ nm}$. Le intensità sono espresse in unità arbitrarie e normalizzate rispetto all'area sottesa alle curve.

relativi, che in genere hanno poca rilevanza pratica essendo associati a piccole frazioni dell'intensità totale, la larghezza del massimo principale dipende fortemente dal rapporto λ/a . Infatti i primi zeri della funzione si trovano nelle posizioni $(\sin \theta)_{min} = \pm \lambda/a$, per cui la “larghezza totale” (misurata tra gli zeri e in unità di $\sin \theta$) del massimo principale è $\sim 2\lambda/a$. Per un fascio di luce perfettamente collimato, come possiamo ipotizzare per il fascio incidente sull'apertura (descritto da un'onda piana, dunque perfettamente collimato), si ha, prima dell'interazione con l'apertura, $\sin \theta = 0$. Il passaggio attraverso l'apertura introduce una divergenza del fascio, che “si allarga” tanto più quanto maggiore è il rapporto λ/a , cioè, a parità di λ , quanto minore è la dimensione dell'apertura a . Ritroverete questa identica affermazione quando affronterete, nell'ambito della meccanica quantistica, l'esperimento ideale che va sotto il nome di *microscopio di Heisenberg*, che consente una spiegazione molto immediata e intuitiva della diffrazione.

Da ultimo, è evidente che anche in questo caso spesso si preferisce nella pratica convertire lo spostamento angolare in spostamento lineare, cosa che si ottiene ponendo lo schermo a distanza $D \gg a$ e misurando la posizione x dei minimi e massimi della figura di diffrazione. Usando le approssimazioni geometriche già ampiamente impiegate, si ottiene che la posizione dei primi minimi si trova a $(x)_{min} = \pm D\lambda/a$. Di conseguenza la misura di x , supponendo noto λ (e D), permette di dedurre quella di a .

A. Diffrazione da apertura circolare

Abbiamo trattato finora delle situazioni sostanzialmente unidimensionali, cioè con aperture “strette e lunghe”, per le quali diffrazione (e interferenza) hanno luogo idealmente solo lungo una direzione cartesiana. Spesso, però,

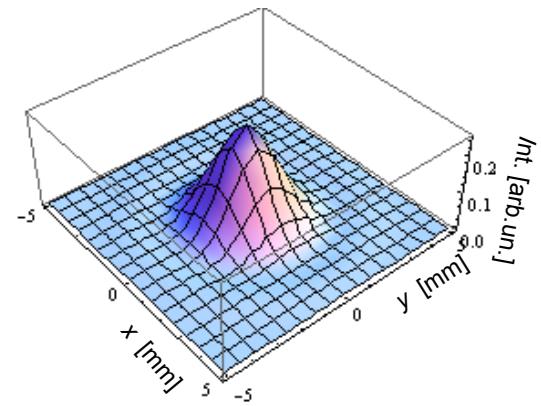


Figura 6. Rappresentazione tridimensionale della figura di diffrazione che si osserva su uno schermo posto a distanza $D = 10 \text{ cm}$ quando radiazione di lunghezza d'onda $\lambda = 500 \text{ nm}$ viene diffratta da un'apertura circolare di diametro $a = 20 \mu\text{m}$. Il calcolo è stato eseguito sulla base dell'Eq. 11.

si ha a che fare con sistemi a simmetria circolare, in particolare con aperture circolari di diametro a . Anche se la fisica dei fenomeni rimane sostanzialmente la stessa, il passaggio da coordinate cartesiane a coordinate circolari (cilindriche) comporta delle differenze di dettaglio, che vale la pena sottolineare.

In particolare, per quello che riguarda la diffrazione, l'uso delle coordinate cilindriche comporta una differente scrittura dell'Eq. 9, che diventa:

$$I(\theta) = I_0 \frac{J_1^2(\alpha)}{\alpha^2} \quad (11)$$

$$\alpha = \frac{a}{\lambda} \sin \theta , \quad (12)$$

dove J_1 rappresenta una (famosa) funzione detta *funzione di Bessel di ordine uno*.

Il primo zero di questa funzione, che fornisce la posizione angolare del primo minimo della figura di diffrazione, si ha quando l'argomento è pari a 1.22, cioè per $(\sin \theta)_{min} = 1.22(\lambda/a)$. Supponendo di usare il solito schermo (posto a distanza $D \gg a$ dall'apertura) e le solite approssimazioni, la figura di diffrazione darà luogo a un sistema di minimi di intensità, cioè di frange, di forma circolare, il cui diametro Φ è legato al diametro a dell'apertura attraverso la relazione, che potete facilmente verificare, $\Phi = 2\pi m D [1.22(\lambda/a)]$, con m intero (spesso detto anche in questo caso ordine di diffrazione). Per intenderci, usando radiazione visibile a $\lambda = 500 \text{ nm}$ e un'apertura (*pin hole*) di diametro $a = 20 \mu\text{m}$, su uno schermo posto a distanza $D = 10 \text{ cm}$ si osserva un primo minimo di intensità che forma un cerchio di diametro $\Phi \approx 6 \text{ mm}$ (la Fig. 6 mostra una rappresentazione tridimensionale di quanto si osserva sullo schermo in questo caso, costruita sulla base dell'Eq. 11).

B. Young revisited

Avevamo preannunciato che saremmo tornati a occuparci dell’interferenza da doppia fenditura (Young) per darne un’interpretazione più realistica. In realtà le due fenditure incise sulla lamina opaca in quell’esperimento hanno una dimensione trasversale finita, che qui indichiamo con a , dunque esse producono diffrazione. Di conseguenza in un esperimento reale, come quello condotto in laboratorio, la figura che si osserva sullo schermo è la convoluzione di due fenomeni: interferenza dal sistema delle due fenditure e diffrazione da ognuna di esse. Infatti nell’esperienza pratica si osservano due distinti sistemi di frange, cioè due distinti sistemi di minimi e massimi di intensità regolari, con spaziature rispettivamente proporzionali a λ/d e λ/a .

La Fig. 7 mostra una simulazione in cui la modulazione dell’intensità dovuta all’interferenza delle due fenditure (Eq. 6) è stata moltiplicata per la modulazione dell’intensità dovuta alla diffrazione da parte delle due fenditure (Eq. 9). In questa ricostruzione numerica si è supposto di osservare le frange di interferenza, ovvero la figura di diffrazione risultante, su uno schermo posto a distanza $D = 1$ m dal piano delle fenditure, per cui l’asse orizzontale rappresenta la posizione sullo schermo. Osservate che l’intensità delle frange di interferenza più lontane dall’asse geometrico del sistema (posizione $x = 0$) è stata moltiplicata per un fattore, come indicato nel grafico, per compensare la forte riduzione di intensità dovuta alla diffrazione. Questa moltiplicazione, che può suonare arbitraria, è in realtà compatibile con la risposta dell’occhio umano, la cui sensibilità è fortemente nonlineare (generalmente quasi logaritmica). Il risultato mostra chiaramente la presenza dei due distinti sistemi di frange.

Ovviamente la diffrazione entra in gioco anche quando si usa nella pratica un reticolo ottico. In questo caso l’effetto è quello di abbattere l’intensità degli ordini di diffrazione superiori, per cui nelle applicazioni spettroscopiche ci si limita spesso a impiegare solo l’ordine $m = \pm 1$.

V. RECIPROCITÀ, “IMPORTANZA” DELLA DIFFRAZIONE, CAMPO LONTANO E CAMPO PROSSIMO

Questo paragrafo conclusivo intende commentare in termini un po’ più generali il fenomeno della diffrazione. È utile in primo luogo ricordare una sorta di principio, generalmente chiamato *principio di reciprocità* (o di Babinet), che fonda la sua esistenza su considerazioni molto generali legate all’invarianza per inversione temporale delle equazioni di Maxwell e dell’equazione d’onda.

Negli esempi che abbiamo discusso, abbiamo sempre considerato che la diffrazione, o l’interferenza, avesse origine dalla sovrapposizione di onde secondarie generate da regioni “vuote” (trasparenti) incise su lamine “opache”. Questo principio stabilisce che si ottengono effetti ana-

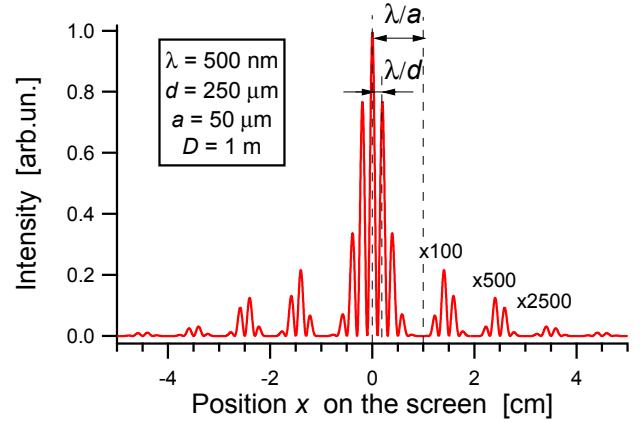


Figura 7. Simulazione dell’intensità su uno schermo (posto a distanza $D = 1$ m) per una radiazione di lunghezza d’onda $\lambda = 500$ nm che incide su un sistema di due fenditure di spaziatura $d = 250$ μm e dimensione trasversale $a = 50$ μm , ottenuta secondo quanto riportato nel testo. In figura sono indicati i fattori moltiplicativi per l’intensità delle frange distanti dall’asse geometrico del sistema (posizione $x = 0$).

loghi, e trattabili con la stessa matematica, se i ruoli di opaco e vuoto si invertono. In altre parole, una struttura materiale opaca inserita in un ambiente trasparente produce diffrazione, sia che abbia una geometria lineare (un sottile filo, un cappello, etc.), sia che abbia una geometria circolare (un granellino di polvere), o di altro tipo.

Dunque la diffrazione può essere davvero considerata come un fenomeno “universale” nel mondo ondulatorio. In ottica essa pone degli argomenti fondamentali (non “tecnologici”) che producono conseguenze rilevanti in tantissimi ambiti. Vediamone qualcuno a parole.

In un qualsiasi esperimento, usare radiazione luminosa implica l’impiego di componenti (lenti, specchi, etc.) che hanno sempre necessariamente delle dimensioni finite. Dunque i fasci hanno dimensioni trasversali finite e per questo motivo la descrizione con onde piane (dove, ricordate, i fronti d’onda sono virtualmente infiniti in direzione trasversale) non è adeguata. La diffrazione introduce un’ulteriore difficoltà: un fascio di luce non può essere mai considerato come completamente collimato, per cui non solo l’estensione trasversale dei fronti d’onda è finita, ma essa, in qualche misura, dipende anche dalla posizione lungo la direzione di propagazione dell’onda.

La diffrazione ha anche un’altra conseguenza fondamentale: a differenza di quanto prevede l’ottica geometrica, un fascio non può neanche essere completamente focalizzato a formare un punto. Qualsiasi sia il sistema ottico (lente, obiettivo, o altro) impiegato per ridurre le dimensioni del fascio, cioè per *focalizzarlo*, le dimensioni trasversali dello spot focale saranno sempre finite.

Sfruttiamo, in una forma un po’ diversa rispetto a quanto enunciato prima, il principio di reciprocità per affermare che le minime dimensioni dello spot sono paragonabili alle minime dimensioni trasversali di un oggetto puntiforme (opaco o trasparente che sia) misurate con un

microscopio ottico convenzionale. Un microscopio ottico è uno strumento che, usando una combinazione di lenti (obiettivo e oculare), è in grado di fornire un'immagine ingrandita di un oggetto. L'ingrandimento dipende solo dalle caratteristiche delle lenti usate e può essere virtualmente reso grande a volontà.

Tuttavia, a causa della diffrazione l'immagine dell'oggetto puntiforme verrà allargata, cioè si formerà una figura di diffrazione con un "diametro" finito e non nullo. Se supponiamo di avere due distinti oggetti puntiformi, il nostro microscopio, per quanto raffinato, ci permetterà di distinguerli solo se le figure di diffrazione prodotte dai due oggetti sono "abbastanza" separate l'una rispetto all'altra. Storicamente è stato introdotto un criterio (detto *criterio di Rayleigh*, peraltro piuttosto ottimista) che ha condotto a un limite, detto *limite di Abbe*, che stabilisce che il massimo *potere risolutivo*, cioè la minima distanza a cui possono essere collocati due oggetti puntiformi per essere apprezzati come distinti, è tipicamente dell'ordine di 0.6λ . Questo limite è in effetti una riscrittura dell'Eq. 11, cioè nasce da una manipolazione matematica di quella equazione, a testimonianza che esso è dovuto alla diffrazione. Dunque, per quanto bravi siate stati nel costruire le lenti del vostro microscopio, potrete usarlo per determinare i dettagli di oggetti che hanno dimensioni minime dell'ordine di 0.6λ , che nel visibile significa diverse centinaia di nm (decisamente troppo per la nanotecnologia).

Se applichiamo di nuovo in senso inverso il principio di reciprocità, possiamo concludere che le dimensioni trasversali minime dello spot focale sono dell'ordine di 0.6λ , anche qui a prescindere dalle qualità dei componenti usati per focalizzare. Nei CD e DVD, che abbiamo già ricordato in precedenza, la scrittura/lettura avviene per via ottica. Il DVD, grazie al pitch minore rispetto al CD, dà la possibilità di raggiungere una densità di immagazzinamento ottico dei dati nettamente superiore. Bene, può essere interessante sapere che l'evoluzione tecnologica dall'uno all'altro supporto ottico è stata soprattutto conseguenza della disponibilità di sorgenti (laser a dio-dio) operanti a lunghezze d'onda minori, da oltre 800 nm per i primi CD, a 660 nm per i DVD. Una delle ultime evoluzioni di questa tecnologia, il Blue-Ray, deve il suo nome al fatto di impiegare sorgenti laser nel blu (405 nm). Infatti, sulla base di quanto abbiamo appena stabilito, ridurre la lunghezza d'onda consente (assieme a tanti altri dettagli di tipo tecnologico) di ridurre lo spot focale, e quindi di aumentare la densità dei "bit ottici" che contengono l'informazione.

Infine facciamo un'ultima considerazione: di fatto, in tutta la nostra trattazione abbiamo seguito un approccio di . In particolare l'andamento dell'intensità per diffrazione che abbiamo ottenuto (*diffrazione di Fraunhofer*) vale solo se la distanza a cui verifichiamo gli effetti della diffrazione stessa è molto maggiore della dimensione dell'apertura (in pratica, $D \gg a$). Vale la pena di ricordare che la diffrazione produce effetti estremamente interessanti anche per distanze molto piccole,

nel regime che si chiama di *campo prossimo*. Interpretare questi effetti richiede di usare una matematica diversa, nella quale, ad esempio, il carattere "propagante" delle onde elettromagnetiche non è più rilevante (se vi ponete a piccola distanza da un insieme di dipoli oscillanti, il ritardo di fase delle varie onde dovuto alla propagazione può diventare trascurabile). È interessante notare che l'opportunità di porsi a piccola distanza dalla sorgente della diffrazione (usare il campo prossimo) è talvolta impiegata proprio per superare i limiti di risoluzione spaziale dovuti alla diffrazione, cioè per costruire microscopi con elevatissimo potere risolutivo o localizzare la radiazione luminosa in regioni spazialmente limitate.

APPENDICE I

In questa Appendice si mostra un possibile procedimento che porta a trovare l'Eq. 7 per la diffrazione da reticolo; questo procedimento ricalca quello discusso da Hecht nel suo testo di Ottica.

Il problema è schematizzato in Fig. 8: essa è simile alla Fig. 3(a), solo che stavolta le aperture, o fenditure, che compaiono in gran numero, sono disegnate più piccole. Notate che anche in questo caso la condizione $D \gg d$ non è rispecchiata nella figura.

Numeriamo le fenditure, che sono in totale N , con l'indice n che corre da 0 a $N - 1$. L'ampiezza del campo elettrico E sullo schermo sarà data dalla parte reale della seguente espressione, in cui sono sommate tutte le onde che interferiscono tra loro:

$$E = \sum_{n=0}^{N-1} E_0 \exp[i(\vec{k}_n \cdot \vec{r}_n - \omega t)] = \quad (13)$$

$$= E_0 \exp(-i\omega t) \sum_{n=0}^{N-1} \exp(ikr_n), \quad (14)$$

dove \vec{r}_n è il vettore che congiunge il (centro) dell'apertura n -esima con il punto di osservazione con lo schermo e \vec{k}_n è il vettore d'onda del "raggio" corrispondente. Per costruzione è $\vec{k}_n // \vec{r}_n$ e, inoltre, $|\vec{k}_n| = k = 2\pi/\lambda$, da cui i passaggi eseguiti. Osservate che, se $D \gg d$, tutti i vettori \vec{k}_n e tutti i "raggi" \vec{r}_n tendono a essere paralleli tra loro. Nell'Eq. 13 abbiamo poi messo in evidenza l'ampiezza E_0 dell'onda diffratta dalle fenditure e il termine di oscillazione temporale, in modo da poterci concentrare sulla serie, che è l'aspetto di interesse per il calcolo.

Mettiamo anche in evidenza il termine $\exp(ikr_0)$ che corrisponde all'onda prodotta (ovvero diffratta) dalla fenditura marcata con $n = 0$:

$$\sum_{n=0}^{N-1} \exp(ikr_n) = \exp(ikr_0) \sum_{n=0}^{N-1} \exp[ik(r_n - r_0)]. \quad (15)$$

Ragionando in termini simili a quanto fatto nella discussione dell'interferenza da doppia fenditura (Young), possiamo porre per la differenza di cammino ottico

$$r_n - r_0 \simeq n\delta, \quad (16)$$

dove δ , differenza di cammino ottico tra i raggi uscenti dalle fenditure 1 e 0, è definita in analogia con Fig. 3(a).

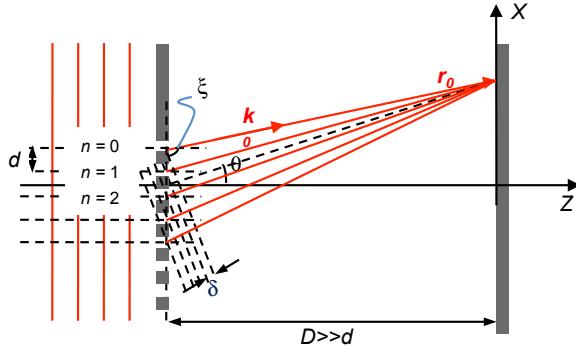


Figura 8. Schema del reticolo considerato in Appendice I, con indicate le grandezze rilevanti per il calcolo: notate la condizione $D \gg d$, che non può essere rappresentata adeguatamente in figura.

La serie diventa allora

$$\sum_{n=0}^{N-1} \exp(ikn\delta) = \frac{1 - \exp(iNk\delta)}{1 - \exp(ik\delta)}, \quad (17)$$

dove l'ultimo passaggio sfrutta l'espressione della somma parziale che si usa nel calcolo delle serie geometriche.

Facciamo ancora un po' di maquillage:

$$\frac{1 - \exp(iNk\delta)}{1 - \exp(ik\delta)} = \frac{\exp(-ikN\delta/2)}{\exp(-ik\delta/2)} \times \quad (18)$$

$$\times \frac{\exp(ikN\delta/2) - \exp(-ikN\delta/2)}{\exp(ik\delta/2) - \exp(-ik\delta/2)}. \quad (19)$$

Ricordiamoci ora che quello che si osserva sullo schermo è l'*intensità* dell'onda ottenuta per sovrapposizione, che è proporzionale al modulo quadro dell'ampiezza del campo. Tenendo conto che tutti gli esponenziali con argomento immaginario che si trovano a moltiplicare hanno modulo unitario (inclusi quelli che avevamo messo in evidenza in Eqs. 13, 15), avremo che l'intensità sarà

$$I \propto \left| \frac{\exp(ikN\delta/2) - \exp(-ikN\delta/2)}{\exp(ik\delta/2) - \exp(-ik\delta/2)} \right|^2 = \frac{\sin^2(kN\delta/2)}{\sin^2(k\delta/2)}. \quad (20)$$

Infine, notando in analogia con la discussione svolta in Sect. III A che $\delta \simeq d \sin \xi \simeq d \sin \theta$, dove le approssimazioni sono tante più valide quanto più D è maggiore di d , e ricordando che $k = 2\pi/\lambda$, si ottiene

$$I(\theta) \propto \frac{\sin^2(N\pi d \sin \theta / \lambda)}{\sin^2(\pi d \sin \theta / \lambda)}, \quad (21)$$

che è quanto annunciato in Eq. 7, dove si era posto $\gamma = \pi d \sin \theta / \lambda$.

APPENDICE II

In questa Appendice si riporta un metodo per la determinazione dell'Eq. 9 per la diffrazione da singola fenditura lineare. Stavolta il procedimento non è del tutto convenzionale e fa uso di qualche shortcut.

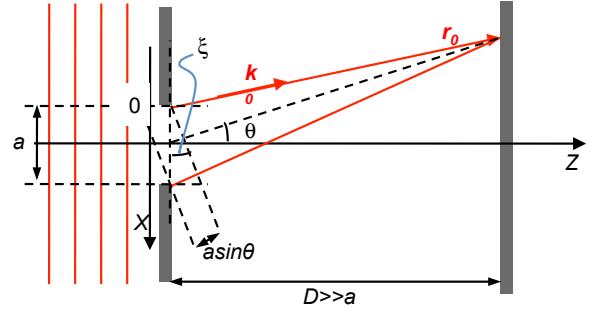


Figura 9. Schema della singola fenditura considerata in Appendice II, con indicate le grandezze rilevanti per il calcolo: notate la condizione $D \gg a$, che non può essere rappresentata adeguatamente in figura.

L'idea di fondo, già annunciata in Sezione IV, è quella di suddividere la fenditura, che ha apertura a , in tanti (infiniti) intervallini molto piccoli (infinitesimi) ai quali applicare il metodo usato per calcolare la diffrazione dal reticolo presentato in Appendice I. Idealmente, quindi, si ha a che fare con un numero $N \rightarrow \infty$ di piccole (infinitesime) aperture la cui spaziatura d tende a zero, mantenendo $Nd = a$.

Suddividiamo allora la fenditura in tanti elementini di lunghezza infinitesima dx (l'asse X corre sul piano dell'apertura e ha origine nell'estremo "alto" dell'apertura stessa, come in Fig. 9). A ognuno di questi elementini, che si comporterà da sorgente secondo il principio di Huygens, associamo una "densità lineare" di ampiezza di onda pari a E_0/a .

Ragionando in maniera simile a quanto fatto in Appendice I, potremo scrivere l'*ampiezza* (in forma complessa) del campo elettrico su un punto dello schermo, che dista $D \gg a$ dal piano della fenditura, sotto forma di *integrale*:

$$E = \int_0^a \frac{E_0}{a} \exp[i(\vec{k} \cdot \vec{r} - \omega t)] dx = \quad (22)$$

$$= E_0 \exp(-i\omega t) \exp(ikr_0) \int_0^a \frac{\exp[ik(r - r_0)]}{a} dx \quad (23)$$

dove la simbologia, costruita in analogia con Eq. 13, dovrebbe essere autoesplicativa così come dovrebbero risultare chiari i vari passaggi.

Occupiamoci del solo integrale, che rappresenta l'aspetto di interesse. facciamo un cambio di variabile passando alla variabile $\zeta = (r - r_0)$, che rappresenta la differenza di cammino ottico. Essa dipende dalla posizione x lungo la fenditura: usando tutte le approssimazioni già ampiamente discusse, si ottiene $\zeta \simeq x \sin \xi \simeq x \sin \theta$, dove l'angolo θ è definito in analogia con le derivazioni discusse in precedenza. Si ha quindi $dx = d\zeta / \sin \theta$; inoltre gli estremi di integrazione diventano 0 e $a \sin \theta$. In definitiva si ha

$$\int_0^{a \sin \theta} \frac{\exp[ik(r - r_0)]}{a} dx = \quad (24)$$

$$= \int_0^{a \sin \theta} \frac{\exp(ik\zeta)}{a \sin \theta} d\zeta = \quad (25)$$

$$= \frac{\exp(ika \sin \theta) - 1}{ika \sin \theta} = \quad (26)$$

$$= \exp(ika \sin \theta/2) \times \quad (27)$$

$$\times \frac{\exp(ika \sin \theta/2) - \exp(-ika \sin \theta/2)}{ika \sin \theta}, \quad (28)$$

dove nei vari passaggi abbiamo calcolato l'integrale e fatto qualche altra piccola manipolazione matematica.

Ricordiamo che ci interessa determinare l'*intensità* dell'onda ottenuta per sovrapposizione sullo schermo. Dobbiamo quindi considerare l'andamento del modulo quadro del campo elettrico. Tenendo conto che i termini a

modulo unitario messi in evidenza nelle Eqs. 22, 24 non hanno alcun ruolo, si ottiene facilmente

$$I(\theta) \propto \left| \frac{\exp(ika \sin \theta/2) - \exp(-ika \sin \theta/2)}{2ka \sin \theta/2} \right|^2 = \quad (29)$$

$$= \frac{\sin^2(ka \sin \theta/2)}{(ka \sin \theta/2)^2}. \quad (30)$$

Considerando che, al solito, $k = 2\pi/\lambda$, si trova infine

$$I(\theta) \propto \frac{\sin^2(\pi a \sin \theta/\lambda)}{(\pi a \sin \theta/\lambda)^2}, \quad (31)$$

che è quanto annunciato in Eq. 9, dove si era posto $\alpha = \pi a \sin \theta/\lambda$.

[1] Il cammino ottico è il prodotto tra distanza percorsa e indice di rifrazione del mezzo in cui l'onda si propaga.

Qui supponiamo di essere nel vuoto, per cui cammino e cammino ottico coincidono.

Laser, laser a diodo, fotodiodo a livello zero

francesco.fuso@unipi.it; <http://www.df.unipi.it/~fuso/dida>

(Dated: version 2 - FF, 16 maggio 2016)

Il laser può essere considerato come una delle “applicazioni” più eclatanti della meccanica quantistica. Descriverne il funzionamento richiede quindi di usare concetti e tecniche che non sono alla portata degli studenti del secondo anno di fisica. Trattare poi di laser a diodo e fotodiode pone l’ulteriore problema di modellare il comportamento dei materiali, che a sua volta richiede concetti piuttosto avanzati e specifici. In questa breve nota si cerca di trattare gli aspetti fondamentali coinvolti nell’operazione di un laser, con particolare riferimento al laser a diodo (e con cenni ai fotodiodi), mantenendo al minimo, a “livello zero”, l’introduzione di nuovi concetti. Sarete in grado di revisionare criticamente tutto quanto qui riportato proseguendo nei vostri studi.

I. LASER

L’invenzione del laser, avvenuta una cinquantina di anni fa grazie alla straordinaria e fruttuosa convergenza di approcci di ricerca industriale e accademica in tante parti del mondo, ha aperto la strada a un’incredibile varietà di sviluppi applicativi e fondamentali, che sono tuttora un vivissimo argomento di ricerca in fisica.

Il termine laser è una sigla, presto diventata un acronimo. Laser significa *Light Amplifier Stimulated Emission Radiation*. In questa sigla è contenuta la gran parte delle caratteristiche di funzionamento del laser, che è appunto un dispositivo in cui si verifica *amplificazione* della radiazione luminosa (tipicamente nel visibile o nel vicino infrarosso) grazie a processi di *emissione stimolata*. La sigla, però, omette di specificare che nel laser l’amplificazione di radiazione è normalmente accoppiata a fenomeni di retroazione (feedback) che portano il dispositivo a operare come un oscillatore, cioè a produrre, non solo amplificare, radiazione luminosa in modo “autonomo”.

Le caratteristiche speciali della luce emessa da un laser sono ben note e possono essere riassunte affermando che il laser emette radiazione *coerente*. L’elevata coerenza della radiazione, che può essere declinata in varie forme (spettrale, temporale, spaziale), è alla base di quelle proprietà che non possono essere riprodotte con altre sorgenti (lampade, per esempio), come la possibilità di essere focalizzata in modo efficace (al limite di diffrazione), l’eccellente livello di monocromaticità, la capacità di essere impiegata in misure interferometriche di tante diverse tipologie, e altre che qui non citiamo.

Dal punto di vista corpuscolare, la caratteristica di coerenza implica che il fascio di un laser possa essere qualitativamente descritto da un fascio di fotoni che hanno *tutti* le stesse proprietà (energia, quantità di moto e anche momento angolare). È evidente che un fascio così realizzato permette di disporre di uno strumento che, una volta messi in atto opportuni di schemi di interazione con la materia, realizza un grado di controllo elevatissimo sulla materia stessa. Per esempio, in conseguenza di questo elevato grado di controllo, manipolazioni e analisi della materia condotte con luce laser possono portare a modifiche fortemente selettive nella materia, oppure a misure

estremamente sensibili delle grandezze analizzate. Anche se, come in parte discuteremo nel seguito, in un laser reale la coerenza non può essere completa, ugualmente l’uso del laser porta a una generalizzata capacità di controllare i sistemi materiali che non può essere ottenuta con altre sorgenti. Questo è uno dei principali motivi per l’importanza e la diffusione dei laser nelle applicazioni scientifiche degli ultimi decenni.

A. Componenti del laser

Per scopi “didattici”, è possibile distinguere diverse componenti, materiali o concettuali, che concorrono al funzionamento di un laser:

1. il *mezzo attivo*, che è un sistema materiale che, attraverso interazione con la luce, permette di ottenere amplificazione;
2. il *pompaggio*, che rappresenta l’insieme di tecniche, pratiche e concettuali, che rendono possibile l’amplificazione di luce da parte del mezzo attivo;
3. la *cavità ottica*, che permette di selezionare o controllare le caratteristiche della luce laser e, soprattutto, di modificare il funzionamento del mezzo attivo da amplificatore a oscillatore, necessario affinché il laser possa operare come sorgente di radiazione.

II. MEZZO ATTIVO

Alla base del funzionamento di qualsiasi laser c’è la realizzazione di specifici schemi di interazione tra radiazione (nel range di interesse per l’ottica, quindi luce, in senso estensivo) e materia. L’interazione radiazione-materia è un argomento estremamente vasto, all’interno del quale convergono moltissimi modelli sviluppati per interpretare diversi fenomeni. Gran parte di questi modelli impiega in maniera pesante descrizioni di tipo quantistico.

Non è questa la sede opportuna per affrontare l’argomento; tuttavia, è utile e necessario richiamare qualche

concetto molto generale, e a livello assolutamente qualitativo. In particolare occorre chiarire che, in un contesto specifico di semplificazioni modellistiche e approssimazioni, la materia di nostro interesse dispone di *livelli discreti di energia*. Tenerne conto aiuta in maniera sostanziale la comprensione del funzionamento dei laser.

Se prendiamo l'esempio più semplice di "materia" rilevante in questo ambito, un atomo di idrogeno descritto dal modello semiclassico di Bohr, la presenza di livelli discreti di energia è evidente. Basta infatti applicare la regola di quantizzazione del momento angolare a diversi valori del numero quantico principale n per ottenere orbite di raggio diverso. A queste orbite corrispondono specifici valori di energia cinetica e di energia di interazione elettrostatica, e dunque compaiono livelli di energia complessiva diversa (tutti negativi e crescenti con n), come mostrato in Fig. 1(a). Notate che, idealmente e nell'ambito del nostro modello, questi livelli di energia sono perfettamente definiti, cioè l'energia di ogni livello, ovvero la differenza di energia fra diversi livelli, può essere data senza alcuna incertezza.

Naturalmente il modello di Bohr rappresenta una descrizione molto approssimata che può essere applicata sensatamente solo nel caso degli idrogenoidi (e, possibilmente, per alti valori di n). Modelli più sofisticati, in cui l'aspetto quantistico diventa via via più rilevante, valgono per sistemi di maggiore complessità, come atomi non idrogenoidi, molecole, fino ad arrivare allo stato condensato (liquidi, solidi). È difficile in questo contesto individuare una linea guida che permetta una descrizione unificata, ma, come in gran parte vedrete nel prosieguo dei vostri studi, è in genere possibile affermare che, mano a mano che il sistema aumenta la sua complessità, cioè aumenta il numero di componenti elementari (atomici) di cui esso è costituito, i livelli di energia del sistema elementare originario tendono a modificarsi. Come rappresentato schematicamente in Fig. 1(b), finché il numero di componenti elementari, indicato con N in figura, è limitato, come per esempio in un sistema molecolare, i livelli si mantengono discreti, ma ogni singolo livello si suddivide in più "sotto-livelli" (*splitting* di energia). Quando invece il numero di componenti è molto grande, come succede per un sistema allo stato condensato (liquido, solido), allora i livelli discreti degenerano in *bande di energia*.

Andando avanti con i vostri studi, vedrete che responsabili delle modifiche allo schema dei livelli energetici sono le interazioni (Coulombiane, ma anche di altro tipo, spesso squisitamente quantistico) fra i singoli componenti elementari del sistema considerato. Per il momento tenete presente che esistono laser che sfruttano mezzi attivi di ogni genere, atomico, molecolare, liquido, solido, e che normalmente la complessità nell'interpretazione del funzionamento aumenta con la complessità del sistema stesso. In altre parole, un laser che sfrutta un mezzo attivo atomico, per esempio un vapore (il laser HeNe ne è un ottimo rappresentante), è molto più semplice da interpretare che non un laser allo stato solido (il laser a dio-dio ricade purtroppo in questa tipologia). Dunque finché

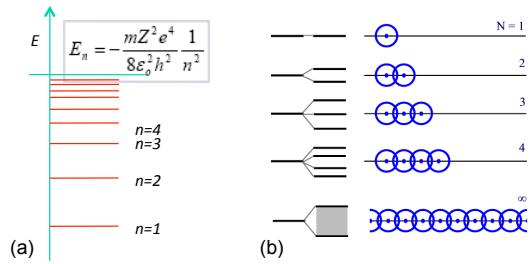


Figura 1. Rappresentazione schematica dei livelli discreti di energia in un atomo di Bohr con numero quantico n (a) e illustrazione dei livelli di energia e della loro origine in sistemi complessi (b), costituiti da un numero N di elementi via via crescente. Illustrazione tratta dal web.

possibile faremo riferimento a casi estremamente semplificati, in cui la materia è costituita da atomi non interagenti fra loro e dotati di livelli di energia discreti.

A. Interazione radiazione materia alla Einstein

Facciamo interagire un "atomo quantistico" con la radiazione. Per comodità, o necessità, decidiamo di descrivere anche la radiazione in maniera quantistica: dunque l'evento elementare è costituito dall'interazione di *un* fotone con *un* atomo. Volendo, nell'ottica di strassemplificare il linguaggio usato, potete pensare a questo evento elementare come a un "urto" (un fotone che arriva sull'atomo) o a una "frammentazione" (da un atomo emerge un fotone), stando però attenti a ricordare che una tale semplificazione non rende onore alla quantità e qualità dei processi effettivamente coinvolti.

La descrizione di Bohr contiene già al suo interno un meccanismo in cui compaiono fotoni: questo meccanismo è il decadimento di un atomo da un livello eccitato a un livello meno eccitato (o al fondamentale). Questo processo rappresenta l'*emissione spontanea* da parte di un atomo eccitato. In esso, come in tutti i processi che qui citeremo, devono essere conservate delle grandezze complessive del sistema atomo/fotone. In particolare, occorre che si conservi l'*energia totale*. Dunque il fotone prodotto dal decadimento deve avere un'energia $E_{phot} = h\nu = E_{n''} - E_{n'}$, dove $E_{n''}$ e $E_{n'}$ sono i livelli di energia di partenza e di arrivo nel processo; la Fig. 2(a) mostra una visione pittorica del processo considerato. È interessante osservare che negli atomi idrogenoidi (più in generale negli atomi) queste differenze di energia cadono proprio nella regione del visibile, o nei suoi dintorni.

Oltre all'emissione spontanea, possiamo facilmente individuare un altro processo che ha solide basi qualitative (esso può anche essere identificato con una trattazione puramente classica): se un fotone arriva su un atomo, è possibile che esso venga assorbito [vedi Fig. 2(b)], cioè che la sua energia venga trasferita al "sistema", come in una sorta di urto anelastico. Per la conservazione dell'energia, il processo di *assorbimento* richiede che il foto-

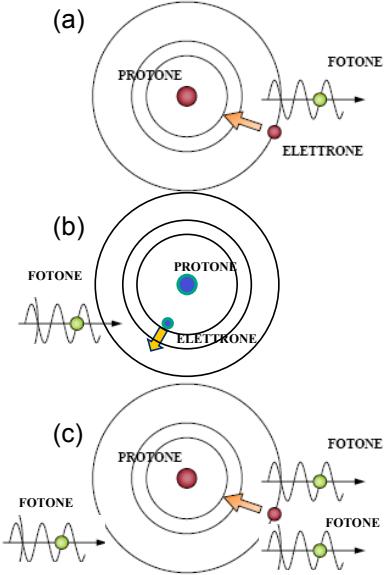


Figura 2. Rappresentazione pittorica dei processi di emissione spontanea (a), assorbimento (b), emissione stimolata (c) a carico di un atomo di Bohr. Illustrazione tratta dal web.

ne abbia un'energia, ovvero la luce abbia una lunghezza d'onda, ovvero una frequenza, *risonante* con la differenza di energia tra due livelli energetici. In questo caso il fotone può “scomparire” e l'elettrone del nostro atomo di Bohr essere promosso a un livello maggiormente eccitato.

Secondo Einstein, che si occupò del problema in uno dei suoi famosi lavori di inizio '900, accanto a questi due processi occorre ipotizzarne un terzo. Questo processo si può verificare a carico di un atomo *che si trova “già” a un livello eccitato* e prevede che, in seguito all'arrivo di un fotone (come al solito di energia risonante), avvenga una diseccitazione accompagnata da emissione di un (altro) fotone [vedi Fig. 2(c)]. A questo processo, che non ha un semplice analogo classico, si dà il nome di *emissione stimolata*. Sulla base di ragionamenti molto semplici, Einstein riuscì a determinare la probabilità con cui avvengono i tre distinti processi, e a legarle alle caratteristiche della materia e della radiazione.

È importante ricordare che la meccanica quantistica “seria” richiede precisazioni e impone limitazioni ai processi che abbiamo considerato, la cui interpretazione dettagliata è tutt’altro che banale. Coerentemente con gli scopi di questa nota, trascuriamo completamente questi aspetti.

B. Emissione stimolata

Due osservazioni importantissime sul meccanismo dell'emissione stimolata:

- il fotone che incide sull'atomo *non* viene assorbito, ma serve solo per “triggerare” il processo: esso si ritrova quindi “in uscita” assieme al fotone

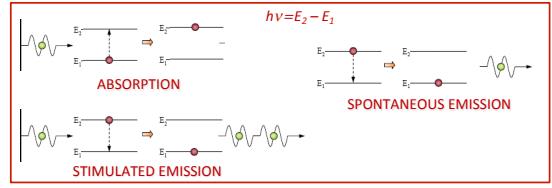


Figura 3. Ulteriore rappresentazione schematica dei processi di assorbimento, emissione stimolata, emissione spontanea. Illustrazione tratta dal web.

che è prodotto dalla transizione dell'atomo a livello energetico inferiore;

- per ragioni di tipo fondamentale, i fotoni emessi per emissione stimolata sono *indistinguibili* dai fotoni che triggerano il processo; notate che, invece, i fotoni emessi per emissione spontanea, pur avendo energia definita (e quindi corrispondendo tutti a luce di una certa frequenza o lunghezza d'onda), hanno quantità di moto, ovvero direzione di emissione, distribuita in modo virtualmente casuale. Essi infatti possono essere considerati emergere in modo spazialmente isotropo dall'atomo che li produce.

A questo punto è molto semplice capire perché la sigla laser enfatizza l'emissione stimolata: questo processo è evidentemente in grado di aumentare il numero di fotoni, dunque *amplificare* la radiazione, e i fotoni emessi sono tutti simili tra loro, cioè l'emissione è *coerente*.

La Fig. 3 mostra un nuovo quadro riassuntivo dei tre processi considerati, stavolta rappresentati ponendo l'accento sulla presenza di livelli energetici.

III. POMPAGGIO E INVERSIONE DI POPOLAZIONE

La possibilità di amplificare radiazione attraverso meccanismi di emissione stimolata potrebbe suonare contraria alle esigenze generali di bilancio energetico. Così ovviamente non è. Infatti per avere emissione stimolata occorre che l'atomo si trovi “già” a uno stato eccitato, cosa che comporta un certo dispendio energetico a monte del meccanismo.

Invece di un singolo elemento (per esempio, atomo) in grado di interagire con la radiazione, consideriamo ora un sistema fatto di tantissimi elementi, cioè un *campione* di elementi. Tantissimi, infatti, sono i fotoni che vogliamo produrre, per cui non è sufficiente considerare un singolo elemento, o atomo, come coinvolto nel processo. Il campione di cui parliamo è generalmente caratterizzato da una distribuzione statistica delle sue grandezze, un po' come succede per gli atomi o le molecole di un gas. Senza entrare nei dettagli della meccanica statistica, è convinzione generalizzata che il campione tenda ad assumere configurazioni in cui l'energia è minimizzata rispetto alle condizioni (in particolare la temperatura) in cui esso si

trova. In altre parole, in un mezzo attivo lasciato evolvere spontaneamente a una data temperatura, troveremo che la maggior parte dei suoi elementi sta a livello energetico basso (semplificando, il fondamentale). È importante ricordare che i livelli energetici di cui ci stiamo occupando hanno separazioni che corrispondono a fotoni nel visibile (o dintorni), cioè nel range dell'eV. La temperatura ambiente, invece, corrisponde a un'energia $k_B T \simeq 1/40$ eV, per cui a temperatura ambiente (ma anche a temperature superiori, purché tali da non distruggere la struttura del campione) sarà molto poco probabile avere elementi eccitati nel campione.

Si chiama *popolazione* il numero di elementi di un campione che si trova a un certo livello energetico. Spontaneamente un sistema evolve in modo che la sua popolazione sia quasi tutta allo stato fondamentale. Se si vuole invece che la maggior parte degli elementi del campione si trovi a livelli eccitati, come necessario per l'emissione stimolata, allora occorre realizzare condizioni dette di *inversione di popolazione*. I processi che conducono all'inversione di popolazione si chiamano processi di *pompaggio*.

A seconda della tipologia di laser sono stati messi a punto nel corso del tempo tanti metodi di pompaggio. I principali sono quelli che sfruttano della radiazione luminosa come pompa (pompaggio ottico), che per altro sono anche di gran lunga i più semplici da descrivere, e quelli in cui invece l'inversione di popolazione è realizzata grazie al passaggio di corrente, o alla formazione di scariche elettriche. Nel laser a diodo il pompaggio avviene elettricamente, in seguito all'iniezione di cariche elettriche attraverso la giunzione operata tramite opportuna polarizzazione della giunzione stessa. È ovvio che questo passaggio di corrente richiede della potenza, per cui, alla fine, le considerazioni di bilancio energetico sono ben soddisfatte (anzi, normalmente i laser sono macchine poco efficienti, nel senso che molta dell'energia fornita viene impiegata per fare altro rispetto alla produzione di luce laser). Per rimanere nell'ambito di oggetti piccoli e molto diffusi, il puntatore laser verde, quello che si usa nelle partite di calcio per acceccare i portieri quando devono parare un rigore, è invece un sistema a pompaggio ottico, essendoci al suo interno un ulteriore laser, normalmente a diodo, che emette radiazione (nel vicino infrarosso) che serve a pompare otticamente uno specifico mezzo attivo.

IV. CAVITÀ OTTICA

Sulla base di quanto sopra descritto, si capisce come un mezzo attivo opportunamente pompato si possa comportare da amplificatore di radiazione, dato che al suo interno sono possibili, e anzi prevalenti, meccanismi elementari di emissione stimolata. Però, come già abbiamo affermato, un laser non è propriamente un amplificatore, ma piuttosto un oscillatore in grado di produrre radiazione coerente, e non solo di amplificarla. Nel passaggio da amplificatore a oscillatore ha un ruolo importantissimo la

cavità ottica all'interno della quale viene posto il mezzo attivo opportunamente pompato.

Cominciamo con il chiarire cosa intendiamo per cavità ottica facendo riferimento al sistema più semplice possibile: una cavità costituita da due specchi piani perfettamente riflettenti, posti parallelamente a una certa distanza d l'uno dall'altro, come in Fig. 4. Supponiamo che all'interno di questa cavità, che contiene il mezzo attivo *pompato*, sia presente della radiazione elettromagnetica di lunghezza d'onda λ , per esempio con vettore d'onda \vec{k} orientato lungo l'asse geometrico della cavità (asse Z) e diretto verso la destra di figura. La riflessione dallo specchio di destra provoca la sovrapposizione di due onde contropropaganti, ovviamente della stessa frequenza, che dà luogo a una sorta di interferenza. Si forma un'onda stazionaria che può essere descritta dalla funzione $\vec{E} = E_0 \sin(kz) \cos(\omega t) \hat{e}_x$, dove tutti i simboli dovrebbero avere significato ovvio e dove si è supposto di prendere lo zero dei tempi in modo da esprimere l'andamento temporale come indicato dall'espressione.

La presenza degli specchi impone delle specifiche condizioni al contorno sui campi, che devono essere nulli sulle superfici degli specchi stessi (e per continuità al loro interno). Supponendo che gli specchi si trovino nelle posizioni $z = 0$ e $z = d$, deve essere *in ogni istante* $\sin(kd) = 0$, ovvero $kd = (2\pi/\lambda)d = m\pi$, con m intero [1]: la cavità che stiamo considerando supporta dei modi stazionari di radiazione con lunghezza d'onda λ tale che $d = m\lambda/2$, con m intero.

È facile stabilire quanto vale la differenza *in frequenza* tra due modi (longitudinali) supportati, per esempio corrispondenti alle lunghezze d'onda λ_1 e λ_2 . Ponendo $\lambda_{1,2} = 2d/m_{1,2}$, con m_1 e m_2 interi consecutivi [tali, cioè, che $(m_2 - m_1) = 1$], si ha $\Delta\nu_{fsr} = (c/n)/\lambda_1 - (c/n)\lambda_2 = c/(2nd)$, con n stavolta indice di rifrazione del mezzo attivo (supposto riempire il volume della cavità). A questa separazione in frequenza si dà spesso il nome di *free spectral range*.

Immaginate ora di avere a un dato istante un fotone di energia tale che la lunghezza d'onda della radiazione corrispondente soddisfi la condizione appena detta. Supponiamo poi che il fotone si trovi nella cavità con quantità di moto diretta lungo l'asse della cavità stessa (asse Z): questo fotone potrà rimanere intrappolato nella cavità rimbalzando continuamente tra i due specchi. Nel suo rimbalzare, si troverà ad attraversare continuamente il mezzo attivo. Supponendo che questo sia pompato, il fotone potrà triggerare tanti eventi di emissione stimolata, producendo tantissimi fotoni tutti virtualmente uguali a lui, e dunque in grado a loro volta di triggerare ancora eventi di emissione stimolata. In altre parole, la presenza del fotone nella cavità dà luogo a una sorta di processo a "valanga", risultante, dopo pochi "rimbalzi" dei fotoni sugli specchi, in un'enorme amplificazione di radiazione.

Prima di procedere con le debite considerazioni, facciamo due osservazioni di dettaglio: (i) la cavità potrebbe anche essere realizzata con degli specchi dielettrici (le equazioni di Fresnel consentono di determinare la riflet-

tività all’interfaccia fra dielettrici di diverso indice di rifrazione, ovvero fra un dielettrico e il vuoto), portando a condizioni al contorno concettualmente diverse, ma con esiti del tutto simili a quanto affermato per gli specchi metallici; (ii) nel nostro approccio semplificato trascuriamo, almeno finché possibile, il comportamento della cavità in direzione trasversale.

A. Perdite e innescos

Nel modello che stiamo impiegando, avere una cavità con specchi completamente riflettenti implica che il fotone rimbalzi per un tempo *infinito* tra gli specchi. Se il fotone, o i fotoni, restano confinati indefinitamente nella cavità, non possiamo certamente realizzare un laser: questi fotoni, infatti, vogliamo ‘impiegarli’ nel fascio laser, e quindi essi devono in qualche misura uscire fuori dalla cavità.

Infatti, affinché possa essere usata in un laser, la cavità deve avere almeno uno specchio *semiriflettente*, da cui la luce possa uscire (si parla in gergo di *output coupler* proprio per indicare che da questo specchio esce la radiazione laser). L’uscita di una frazione dei fotoni significa che parte dell’energia immagazzinata nella cavità sotto forma di modi stazionari di radiazione viene persa nel tempo. Questo rappresenta un inevitabile meccanismo di *perdita* nell’operazione del sistema che stiamo considerando.

Esistono anche altri inevitabili meccanismi di perdita. Per esempio essi possono essere legati all’assorbimento residuo di radiazione da parte dei materiali (specchi, ma anche mezzo attivo), oppure alla presenza della *difrazione* da parte degli specchi. Infatti essi avranno necessariamente una dimensione finita, e pertanto forniranno alla radiazione una componente non nulla di vettore d’onda in direzione trasversale. In altre parole, potrà verificarsi che alcuni fotoni acquistino quantità di moto in direzione trasversale: dopo aver rimbalzato un certo numero di volte sugli specchi, questi fotoni lasceranno la cavità. Per minimizzare l’effetto della diffrazione è possibile studiare configurazioni ottiche specifiche, per esempio quelle che fanno uso di specchi sferici.

Occupiamoci ora di capire come può essere innescata, cioè “inizidata”, l’emissione laser. Un mezzo attivo pompato ha livelli eccitati popolati. Può dunque esserci emissione spontanea in transizioni da questi livelli eccitati a livelli di energia più bassa. I fotoni prodotti per emissione spontanea possono avere qualsiasi direzione di propagazione, essendo il processo isotropo (nel nostro modello). Potrà sicuramente verificarsi che un fotone (idealmente ne basta uno, trascurando le perdite) sia emesso spontaneamente nella direzione dell’asse della cavità (asse *Z*). Questo solo fotone potrà, se non viene perso, dare origine all’amplificazione “a valanga” che coinvolge l’emissione stimolata.

È evidente che il meccanismo di rimbalzo tra gli specchi creato dalla cavità è in grado, dopo l’innescio, di provvedere continuamente emissione, cioè di fare del laser un

oscillatore in grado di emettere in modo autonomo radiazione. La presenza di una cavità può essere vista come l’aggiunta di un meccanismo di feedback all’amplificatore: infatti la radiazione amplificata viene continuamente (a parte le perdite) re-iniettata nell’amplificatore stesso, portando a condizioni stazionarie di oscillazione. Come certamente ricorderete, abbiamo già accennato in tutt’altro contesto alla possibilità che un feedback applicato a un amplificatore conduca ad oscillazioni.

B. Curva di guadagno e soglia

Abbiamo già specificato che i processi di interazione radiazione materia che stiamo considerando sono *risonanti*, cioè l’energia dei fotoni deve essere pari alla differenza di energia dei livelli coinvolti. Inoltre, per quanto abbiamo affermato, anche la cavità produce una “selezione” delle frequenze. Infatti i modi (longitudinali) che non sono supportati dalla cavità, cioè quelli che corrispondono a lunghezze d’onda diverse dalla condizione prima specificata (numero intero di semilunghezze d’onda pari alla distanza dagli specchi), vengono in qualche modo “persi” dalla cavità. Queste due circostanze determinano la frequenza, o le frequenze, dell’emissione laser e il suo carattere monocromatico.

Limitiamo la nostra casistica a mezzi attivi pompatisi dotati di bande di energia (per esempio, mezzi attivi solidi), che sono quelli di interesse per il laser a diodo. Poiché i livelli energetici non sono discreti, le transizioni sono possibili per un intervallo di energie (la risonanza “si allarga” [2]). Normalmente l’efficienza del processo di amplificazione per emissione stimolata, cioè il *guadagno* fornito dal mezzo attivo, non è uniforme all’interno di questo intervallo di energia, ma ha una forma a campana a cui si dà il nome di *curva di guadagno*. È possibile avere emissione laser quando la curva di guadagno assume valori superiori alle perdite, per esempio quelle dovute a uno dei motivi menzionati in precedenza. Questo determina una *soglia* nell’operazione del laser, che è possibile superare solo se l’inversione di popolazione, ottenuta tramite pompaggio, supera un certo valore critico (al di sotto di questo valore l’amplificazione non è abbastanza efficiente da superare le perdite).

In queste condizioni la frequenza di emissione laser viene determinata dal modo (longitudinale) di radiazione supportato dalla cavità. Poiché in genere il free spectral range della cavità, che abbiamo definito sopra, è piccolo rispetto alla larghezza della curva di guadagno, un singolo modo della cavità viene selezionato e l’emissione laser ha luogo alla frequenza corrispondente. La proprietà della cavità contribuiscono anche a definire il carattere monocromatico dell’emissione: una cavità “migliore” dal punto di vista ottico (con meno perdite) fornisce in genere un’emissione più monocromatica. Spesso, però, si verifica che più modi della cavità siano coinvolti, in maniera generalmente instabile, nell’emissione laser, che quindi si dice “multimodo”. La Fig. 4(b) mostra un’illustrazione quali-

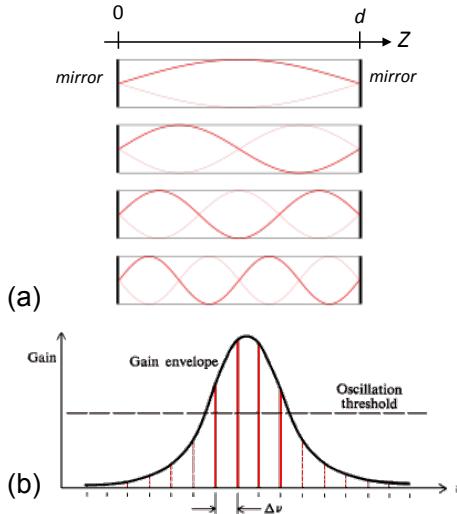


Figura 4. Schema della cavità a specchi piani paralleli (a) e illustrazione di curva di guadagno (gain envelope) e modi della cavità (b): $\Delta\nu$ rappresenta il free spectral range e, nella situazione descritta, sono presenti diversi modi per cui il guadagno supera le perdite (oscillation threshold), cioè l'emissione può virtualmente avvenire a diverse frequenze. Illustrazione tratta dal web.

tativa della curva di guadagno e dei modi della cavità in funzione della frequenza.

V. LASER A DIODO

I primi prototipi, o tentativi di prototipo, di laser a diodo sono coevi all'invenzione dei laser, ma solo negli ultimi decenni questi dispositivi hanno visto un incremento impressionante delle loro caratteristiche e delle possibilità di impiego in un numero crescente di applicazioni di ogni genere. La tecnologia produttiva e la scienza dei materiali hanno avuto un ruolo fondamentale in questi sviluppi (il Nobel Prize 2014 in Fisica è stato attribuito, tra gli altri, all'inventore dei materiali che consentono di produrre laser a diodo con emissione nel blu), rendendo i laser a diodo dispositivi piuttosto sofisticati e complessi.

Qui ci limitiamo a chiarire solo alcuni degli aspetti dell'operazione di un laser a diodo, fornendo, talvolta, informazioni che non corrispondono all'effettiva tecnologia (complicata) dei dispositivi reali.

Per rimanere nell'ambito della descrizione generale dei laser fatta in precedenza, cominciamo con il notare che:

- il mezzo attivo è costituito da una o più *giunzioni* tra materiali semiconduttori drogati;
- il pompaggio avviene tramite passaggio di corrente attraverso la giunzione *polarizzata direttamente*:
- la cavità ottica è "integrata" nel dispositivo, sfruttando in genere la (bassa) riflettività dell'interfaccia semiconduttore/aria ai bordi del dispositivo stesso.

A. Semiconduttori e leghe

Conosciamo già una descrizione qualitativa (classica) dei materiali semiconduttori, con specifico riferimento al Silicio. La descrizione "quantistica" dei semiconduttori, molto complicata da affrontare nei dettagli, non si discosta, in linea di principio, dalle semplicissime considerazioni a cui abbiamo accennato in precedenza in questa nota. I semiconduttori di cui ci occupiamo sono solidi, dunque i loro livelli di energia non sono discreti, ma rappresentati da bande. Gli elettroni che si trovano nella (o nelle) banda di energia più alta, cioè popolano questa (o queste) banda di energia, hanno la possibilità di muoversi all'interno del materiale: la banda di energia più alta si chiama *banda di conduzione*. Gli elettroni che invece si trovano a energia più bassa si dicono popolare la *banda di valenza* e non godono della proprietà di potersi muovere nel materiale.

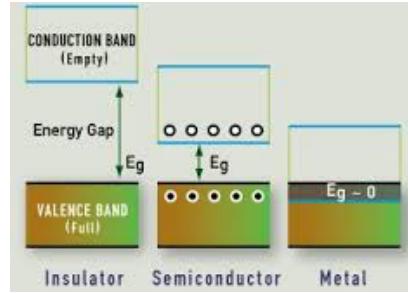


Figura 5. Rappresentazione pittorica delle bande di energia in materiali solidi con diverse proprietà di trasporto elettronico. Illustrazione tratta dal web.

Questa descrizione vale, sempre in linea di principio, per ogni materiale solido. La differenza tra conduttori, semiconduttori e isolanti è rappresentata dalla distanza in energia tra le bande. Come rappresentato in Fig. 5, nei conduttori le bande si sovrappongono, e dunque è sempre possibile avere elettroni in banda di conduzione (gli elettroni liberi). In isolanti e semiconduttori, invece, le bande sono separate da un *energy gap* E_g : negli isolanti il gap è grande (diversi eV) e quindi è molto improbabile, almeno a temperatura ambiente, trovare elettroni liberi. Nei semiconduttori, invece, il gap è sufficientemente piccolo (dell'ordine dell'eV) perché ci sia una probabilità non nulla di avere qualche elettrone in banda di conduzione già a temperatura ambiente. Questi elettroni sono proprio quelli che avevamo identificato come responsabili per la (debole) conducibilità dei semiconduttori intrinseci.

Nella nostra descrizione qualitativa dei semiconduttori avevamo anche identificato dei portatori di carica positiva, le *lacune*. Nella trattazione quantistica a cui stiamo accennando attribuiamo un'energia anche alle lacune e affermiamo, sulla base di ragionamenti qui non riportati, che esse si trovano nella banda di valenza. Infine, avevamo visto come fosse possibile modificare le proprietà di trasporto elettronico di un semiconduttore mediante droggaggio. Il droggaggio altera lo schema dei livelli di energia,

tuttavia tale modifica può essere considerata trascurabile per i nostri scopi. Quindi l'unica variazione di proprietà che riconosciamo al drogaggio consiste nel modificare la popolazione delle bande di energia: il drogaggio n aumenta la popolazione della banda di conduzione, il drogaggio p aumenta quella della banda di valenza.

A questo punto è necessaria una bella precisazione: per motivi molto importanti, che qui non esaminiamo, tra i materiali semiconduttori che si usano per realizzare diodi laser *non c'è il Silicio*. Infatti, nonostante il Silicio sia il paradigma del materiale semiconduttore e sia sicuramente l'ingrediente fondamentale per ogni diodo che non voglia essere un laser, esso non si presta ad essere impiegato in meccanismi di emissione di fotoni, almeno non nel senso convenzionale che è di nostro interesse. Diversi semiconduttori "artificiali", cioè realizzati tramite *leghe* di elementi (spesso appartenenti alle colonne III-V della tavola periodica), rimpiazzano il Silicio per le applicazioni fotoniche, per esempio la lega binaria GaAs (Arseniuro di Gallio) e la sua variante ternaria GaAlAs (Arseniuro di Gallio Alluminio). Nonostante le difficoltà che la fabbricazione di queste leghe comporta, difficoltà che lo sviluppo tecnologico ha comunque permesso di superare brillantemente, la necessità di impiegarle ha diversi risvolti positivi, per esempio quello di ottenere curve di guadagno piuttosto larghe (decine di nanometri, in termini di lunghezza d'onda) e collocate in posizioni spettrali diverse (dal blu/vicino-ultravioletto fino all'infrarosso, anche se con parecchi ed estesi "buchi" in cui non esistono laser a diodo) a seconda dei materiali usati e, in parte, della loro concentrazione nella lega.

B. Giunzione e ricombinazione radiativa

Una giunzione p-n (drogata) polarizzata direttamente sostiene il passaggio di corrente. Questa corrente è costituita da portatori primari (lacune e elettroni nelle regioni rispettivamente p e n). Al passaggio attraverso la giunzione si verificano dei processi di *ricombinazione* in seguito ai quali, per esempio, gli elettroni provenienti dalla regione n si "uniscono" alle lacune presenti nella *zona di svuotamento* (e viceversa le lacune provenienti dalle regioni p si "uniscono" agli elettroni).

La visione quantistica (molto) qualitativa del processo di ricombinazione elettrone-lacuna prevede che in esso l'elettrone passi dalla banda di conduzione a quella di valenza. Nella ricombinazione è quindi coinvolta una variazione di energia, che, nello specifico, deve essere rilasciata. È possibile che questo rilascio di energia avvenga con emissione di un fotone. Più in generale, e a livello decisamente qualitativo, possiamo identificare tra i fenomeni che coinvolgono elettroni e lacune tutti i tre processi (assorbimento, emissione spontanea, emissione stimolata) che abbiamo prima descritto per i sistemi a livelli discreti (atomi, per esempio). Il gioco è allora fatto e siamo in grado di ricondurre il funzionamento delle

giunzioni tra semiconduttori drogati a quanto descritto in precedenza.

C'è inoltre un altro aspetto estremamente rilevante: nelle condizioni che stiamo esaminando (polarizzazione diretta), gli elettroni sono continuamente spostati verso la giunzione dalla d.d.p. applicata tra anodo e catodo del diodo. Dunque essi possono entrare nella zona di svuotamento avendo una densità maggiore di quella delle lacune che vi si trovano. Di conseguenza si possono realizzare in maniera "automatica" le condizioni di inversione di popolazione necessarie per pompare il mezzo attivo, cioè la giunzione polarizzata direttamente si può comportare automaticamente come un mezzo attivo che ha un guadagno molto elevato.

A questo punto sarebbe necessario introdurre una lunga serie di precisazioni e di caveat, che, fortunatamente, vanno al di là degli scopi di questa nota. Qualcosa, però, va accennato:

- l'esito della ricombinazione può condurre alla produzione di fotoni solo se le funzioni d'onda di elettrone e lacuna si sovrappongono spazialmente. Questa affermazione, che abbiamo dato con il linguaggio della meccanica quantistica, si traduce in pratica nel requisito che elettroni e lacune si trovino "vicini" tra loro prima di ricombinarsi. In una giunzione ordinaria, fatta di un solo materiale, la condizione richiesta si realizza con poca probabilità. È possibile aumentare l'efficienza dei processi radiativi di ricombinazione (quelli che coinvolgono fotoni) creando delle giunzioni fra materiali diversi (dette *eterogiunzioni*, per esempio GaAs e GaAsAl), dove la configurazione dei livelli di energia "forza" elettroni e lacune a trovarsi assieme in una piccola regione di spazio a cavallo della giunzione. Storicamente, l'introduzione delle eterogiunzioni è stata un sostanziale passo in avanti per la realizzazione di dispositivi efficienti e durevoli.

- In determinate condizioni, in particolare quando la sovrapposizione spaziale delle funzioni d'onda di elettrone e lacuna è parecchio marcata, si può verificare la formazione di *exciton*, cioè sistemi artificiali legati costituiti da una lacuna e un elettrone, che si comportano in modo qualitativamente simile a protone e elettrone dell'atomo di idrogeno. L'emissione tra i livelli energetici (discreti) dell'excitone può anche contribuire alla realizzazione di emissione laser.

C. Cavità, costruzione e principali caratteristiche

La produzione di luce da giunzioni di semiconduttori (in particolare GaAs e GaAsAl) polarizzate direttamente è sfruttata non soltanto nei laser, ma anche nei LEDs (Light Emitting Diodes). A grandi linee, la differenza principale tra un LED e un laser è nella presenza di una cavità, che consente nel laser di ottenere radiazione da

emissione stimolata. Esistono numerose configurazioni di cavità per laser a diodo, alcune anche molto tricky, ma la più diffusa è certamente quella che prevede che gli specchi di fine cavità siano realizzati dall'interfaccia tra semiconduttore e aria.

Per fare un esempio, il GaAs ha, nel rosso, un indice di rifrazione $n \sim 3.7$. La sua interfaccia con l'aria ($n = 1$) produce, a incidenza normale, una riflettività $R \sim 30\%$ (risultato che si trova agevolmente usando le equazioni di Fresnel). Si può intuire facilmente come una cavità così realizzata abbia notevoli perdite. Dal punto di vista del funzionamento, le grandi perdite non rappresentano un problema significativo, visto l'elevato guadagno del mezzo attivo. Tuttavia la scarsa qualità ottica della cavità influisce in vario modo sulle caratteristiche di emissione del laser a diodo.

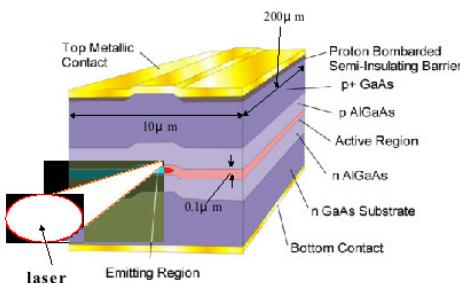


Figura 6. Schema costruttivo di un laser a diodo a eterogunzione della tipologia gain guided. Illustrazione tratta dal web.

La Fig. 6 mostra lo schema realizzativo di un laser a diodo di una tipologia piuttosto comune. Esso si basa su una eterogunzione GaAs/GaAsAl ed è realizzato nella configurazione cosiddetta *gain guided*, in cui l'iniezione di carica attraverso la giunzione avviene in una striscia di dimensioni limitate, definita dalle dimensioni degli elettrodi. Trascuriamo questi dettagli e notiamo qualche aspetto generale nella costruzione del laser a diodo che ha immediati risvolti nelle sue caratteristiche di funzionamento.

- La regione da cui provengono i fotoni ha dimensioni fortemente asimmetriche dovute al fatto che l'emissione avviene solo attraverso la giunzione, grosso modo all'interno della zona di svuotamento che, quando la giunzione è polarizzata direttamente, è spessa poche centinaia di nanometri. Nelle altre due direzioni, invece, le dimensioni sono molto maggiori (anche nel caso di laser gain guided, come in figura). Di conseguenza l'emissione è *asimmetrica e astigmatica*, dando luogo a un fascio non collimato e di forma apparentemente ellittica. Notate che l'asse maggiore dell'ellisse corrisponde alla direzione che attraversa la giunzione: infatti, a causa della diffrazione, la divergenza del fascio è maggiore per la direzione in cui la regione da cui provengono i fotoni è minore. Il forte astigmatismo è in contrasto con la visione (näif) del laser come una sorgente ca-

pace di emettere fotoni tutti con la stessa direzione di propagazione.

- La polarizzazione del fascio laser è legata alla direzione in cui si muovono i portatori di carica. Essi attraversano la giunzione, e quindi la polarizzazione è diretta lungo l'asse maggiore dell'ellisse che descrive il fascio laser.
- La cavità del laser a diodo, oltre a essere di scarsa qualità ottica, è molto corta, tipicamente centinaia di micrometri. Di conseguenza il free spectral range è relativamente grande (decine di GHz). Per condizioni di operazione (corrente e temperatura) adeguate, il laser emette su un singolo modo (longitudinale) della cavità, con una larghezza di riga che può facilmente scendere sotto il MHz (emissione monocromatica). Tuttavia, a causa dell'elevata larghezza della curva di guadagno, è possibile che la lunghezza d'onda cambi, in maniera spesso instabile, da un modo all'altro della cavità.
- L'"efficienza" del laser a diodo è generalmente piuttosto bassa (tipicamente non più del 20-30% della corrente iniettata viene usata per la generazione di luce). Infatti normalmente si usano semiconduttori con droggaggio relativamente basso, quindi i materiali sono caratterizzati da una certa resistività che provoca dissipazione per effetto Joule.
- Di conseguenza, il dispositivo è soggetto a riscaldarsi. Notate che il riscaldamento provoca dilatazione termica della cavità, che comporta una variazione della lunghezza d'onda di emissione. Negli impieghi scientifici, occorre provvedere un sistema di termoregolazione in grado di garantire una stabilità di temperatura notevole (spesso sotto il centesimo di grado centigrado).
- Anche la variazione della corrente iniettata può produrre modifiche della lunghezza d'onda di emissione. Infatti, oltre ad agire indirettamente sulla temperatura, la densità di corrente attraverso la giunzione, che può arrivare a valori molto elevati viste le piccole dimensioni del dispositivo, influenza sull'indice di rifrazione effettivo.
- Naturalmente, affinché ci sia emissione laser, occorre che la corrente iniettata superi un certo valore di soglia. Infatti l'intensità di corrente controlla direttamente l'efficienza di pompaggio e, come in ogni laser, è necessario che il materiale sia pompato al punto che il guadagno superi le perdite (rilevanti, vista la scarsa efficienza della cavità) per avere emissione laser.

In definitiva, le caratteristiche tipiche di un laser a diodo sono poco in accordo con la visione ideale del laser come sorgente in grado di produrre radiazione coerente (per esempio, perfettamente collimata), anche se spesso

caratterizzata comunque da un buon grado di monocromaticità. Tuttavia, oltre a invadere il mercato consumer, negli ultimi decenni i laser a diodo hanno trovato un vastissimo impiego in numerosi esperimenti scientifici, soprattutto grazie alla loro diffusione, economicità, disponibilità in tante varianti.

VI. FOTODIODO

Avendo speso delle parole per descrivere (qualitativamente) il funzionamento del laser a diodo, sarebbe un peccato non servirsene per trattare il funzionamento del fotodiodo, che in un certo senso (molto qualitativamente) è il “negativo” del laser a diodo.

Il fotodiodo è un rivelatore di radiazione, cioè un sensore che fornisce un segnale idealmente proporzionale alla potenza della radiazione che vi incide. Se la radiazione viene descritta in termini di fotoni, allora la potenza rappresenta il numero di fotoni N per unità di tempo, ovvero il loro flusso. Infatti ogni fotone porta una certa energia, $E_{phot} = h\nu$, e il prodotto di questa energia per N fornisce la potenza della radiazione considerata. Tanto per fare un esempio, la potenza $P = 1 \text{ mW}$ di un fascio laser a lunghezza d’onda $\lambda \approx 650 \text{ nm}$ (il laser usato in laboratorio), a cui corrisponde un’energia $E_{phot} = 1240/660 \text{ [eV]} \simeq 1.9 \text{ eV} \simeq 3.0 \times 10^{-19} \text{ J}$, implica un flusso $N = P/E_{phot} \simeq 3.3 \times 10^{15} \text{ fotoni/s}$.

In termini generali, un fotodiodo è un ordinario diodo a giunzione p-n in cui la giunzione è esposta, protetta da una finestra trasparente, alla radiazione. A differenza del laser a diodo, in questo caso *si usa frequentemente il Silicio*, dunque si ha a che fare con qualcosa che costruttivamente è molto simile a un ordinario diodo a giunzione.

Esistono diverse possibilità di operazione per un fotodiodo, a seconda della polarizzazione della giunzione. Qui facciamo riferimento alla cosiddetta modalità *fotovoltaica*, in cui la giunzione è *non polarizzata*. Ricordando quanto affermato nella descrizione delle giunzioni tra semiconduttori drogati, sappiamo che in queste condizioni si forma una regione di svuotamento priva di cariche libere, in cui è presente un campo elettrico detto di *built-in*. Se in questa regione arriva un fotone può verificarsi un processo di *assorbimento*, in seguito al quale l’energia del fotone viene presa dal materiale e un elettrone viene promosso dalla banda di valenza alla banda di conduzione. Contestualmente una lacuna si forma nella banda di valenza. A grandi linee, stiamo parlando di un processo che è opposto a quelli di emissione responsabili per la generazione di fotoni nel laser.

L’elettrone promosso alla banda di conduzione e la lacuna nella banda di valenza si trovano liberi di muoversi sotto l’effetto del campo di built-in. Essi possono quindi raggiungere catodo e anodo del dispositivo e essere raccolti dal circuito esterno di lettura (per esempio, un amperometro), producendo una corrente di intensità misurabile. Il processo è evidentemente lineare con il flusso

di fotoni, cioè con la potenza della radiazione. Quello descritto, in forma molto semplificata, è il meccanismo alla base del funzionamento del fotodiodo in modalità fotovoltaica. La conversione di “energia luminosa” in carica elettrica somiglia all’effetto fotoelettrico ben noto nella storia della fisica, ma rispetto a questo ha un’origine differente.

Qualche considerazione di contorno.

- Affinché l’assorbimento del fotone possa avvenire e dare luogo alla produzione della coppia elettrone-lacuna occorre che la sua energia sia maggiore dell’energia di gap E_g del semiconduttore. Nel caso del Silicio, è $E_g \sim 1.1 \text{ eV}$, per cui è necessario che la radiazione abbia lunghezze d’onda $\lambda < E_g/h \simeq 1.1 \mu\text{m}$.
- Per motivi non semplici da illustrare, l’efficienza del processo dipende dall’energia del fotone: nel caso del Silicio, essa raggiunge il suo massimo per fotoni corrispondenti a lunghezze d’onda nel vicino infrarosso (circa 800 nm).
- In ogni caso, il processo di formazione della coppia elettrone-lacuna ha una certa probabilità di verificarsi. Poiché questo processo compete con altri (per esempio, la “ionizzazione” degli atomi di Silicio dovuti al fatto che il dispositivo si trova a temperatura diversa dallo zero assoluto) e poiché la carica elettrica portata da un singolo elettrone è difficilmente misurabile, il fotodiodo non è adatto per misure di potenze luminose molto deboli, corrispondenti a pochi fotoni al secondo, dove il segnale di corrente è nascosto dal rumore.
- Molto spesso, la costruzione di un fotodiodo presenta delle differenze rispetto a quelle di un diodo ordinario. In particolare, come anche si verifica nei dispositivi in uso in laboratorio (fotodiodi BPW-34), la giunzione pn è rimpiazzata da una giunzione *p-i-n*, in cui la “i” sta ad indicare un sottile (spessore tipico micrometrico) strato di Silicio “intrinseco”, cioè non drogato. Lo scopo principale di questa variante è aumentare lo spessore della giunzione, ovvero della regione di svuotamento nella quale può avvenire il processo di fotogenerazione delle coppie elettrone-lacuna. Questo provoca diverse conseguenze, tra le quali un aumento della “sensibilità” alla radiazione, dovuto all’aumento effettivo del volume della regione “attiva”.

In generale, la modalità fotovoltaica, pur essendo la più semplice tra quelle che possono essere realizzate, soffre di alcuni evidenti limiti. L’estrazione delle cariche dalla giunzione avviene solo per effetto del campo di built-in, poco intenso e poco controllato (la situazione è migliorata nei fotodiodi p-i-n). Inoltre le cariche uscenti dalla giunzione devono attraversare spessori consistenti di materiale semiconduttore prima di giungere a catodo e anodo del dispositivo, dove possono subire effetti dissipativi

dovuti alla resistenza dei materiali stessi. Per superare questi limiti la giunzione può essere polarizzata, sia direttamente che inversamente. In particolare, molto spesso i fotodiodi sono polarizzati inversamente: la polarizzazione inversa aumenta lo spessore della giunzione, consenten-

do una maggiore sensibilità e, soprattutto, un tempo di risposta più rapido. Infatti il tempo di risposta dipende dalla capacità della giunzione, che diminuisce quando lo spessore della giunzione aumenta.

[1] Osservate che la condizione di campo nullo sullo specchio in $z = 0$ è automaticamente soddisfatta dalla scelta della funzione seno per esprimere l'andamento spaziale dell'onda stazionaria. In termini più generali occorrerebbe scegliere una combinazione di seni e coseni, ovvero un seno (o coseno) con un termine di fase costante sommato a kz

nell'argomento.

[2] Per ragioni fondamentali e tecniche, la risonanza non è mai descrivibile con una “funzione delta” dell'energia, o frequenza, ma ci sono sempre processi di allargamento. Però la “larghezza di riga” è sicuramente maggiore per mezzi attivi solidi rispetto a vapori atomici.

Circuiti magnetici e trasformatore

francesco.fuso@unipi.it

(Dated: version 2 - FF, 16 aprile 2021)

Oggetto di questa nota è l'applicazione di alcuni concetti relativi ai circuiti magnetici, in particolare la cosiddetta legge di Hopkinson, la relazione tra induttanza e mutua induttanza con il numero di spire degli avvolgimenti, il funzionamento del trasformatore ideale. Inoltre, allo scopo di meglio contestualizzare l'argomento, una breve premessa serve a richiamare brevemente la tematica dei materiali ferromagnetici e il loro comportamento.

I. MATERIALI FERROMAGNETICI

Come è ben noto, i materiali ferromagnetici hanno spiccate proprietà di carattere magnetico, cioè la loro presenza è in grado di modificare sostanzialmente i campi vettoriali rilevanti per la descrizione dei fenomeni magnetostatici.

Non è certamente questa la sede per entrare nei dettagli di argomenti che fanno ampia parte dei corsi di "Fisica 2", però vale la pena di ricordare che, per motivi pratici, conviene introdurre i tre campi vettoriali \vec{B} , \vec{H} , \vec{M} , regolati dalla relazione $\vec{B} = \mu_0(\vec{H} + \vec{M})$, dove μ_0 è la permeabilità magnetica del vuoto (il valore di questa costante è, nel nostro sistema di unità di misura, $\mu_0 = 4\pi \times 10^{-7}$ T m/A). Il campo di magnetizzazione \vec{M} può essere messo in relazione con la presenza di momenti magnetici \vec{p}_m nel materiale: $\vec{M} = \lim \frac{\Delta N}{\Delta V} < \vec{p}_m >$, dove il passaggio al limite consiste nel considerare volumetti ΔV tendenti a zero e contenenti ΔN momenti di dipolo magnetico, e il bra-ket indica un'operazione di media sull'ensemble, cioè sul sistema di momenti magnetici considerati.

A causa delle loro relazioni costitutive, nel sistema di unità di misura che impieghiamo i tre campi vettoriali non hanno le stesse dimensioni, né le stesse unità di misura. In particolare, l'ampiezza del campo di induzione magnetica \vec{B} si misura in Tesla [T], o l'equivalente Weber/m² [Wb/m²], anche se frequentemente si impiega l'unità alternativa Gauss [G], con l'equivalenza 1 T = 10⁴ G (campi dell'ordine del Tesla sono difficili da ottenere). Vista l'unità di misura di μ_0 , le ampiezze dei campi \vec{M} e \vec{H} risultano misurate in [A/m].

L'utilità dell'impiego dei tre campi vettoriali può essere compresa considerando le equazioni di Maxwell per campi stazionari nella forma $\vec{\nabla} \times \vec{H} = \vec{J}$ e $\vec{\nabla} \times \vec{M} = \vec{J}_m$, dove \vec{J} e \vec{J}_m sono le densità superficiali della corrente fatta da portatori di carica (cariche libere), che in genere scorre intenzionalmente in un qualche conduttore, e della corrente che si assume modelli i fenomeni di magnetizzazione della materia. In un ipotetico esperimento, o situazione fisica, possiamo quindi supporre che \vec{H} sia determinato da correnti note e ben controllate, mentre \vec{M} ha a che fare con la risposta del sistema materiale considerato all'applicazione di \vec{H} , ovvero tiene conto delle proprietà magnetiche intrinseche del materiale.

Trattare questi fenomeni dal punto di vista microscopico è generalmente molto complicato. In primo luogo, l'aggettivo "microscopico" assume in questo ambito un significato particolare, dato che, per esempio, il volumetto ΔV introdotto sopra non può, nella realtà, essere considerato piccolo a piacere essendo limitato dalle scale dimensionali tipiche dei fenomeni che si stanno considerando. Per semplificare la trattazione, si possono poi fare alcune approssimazioni, consistenti nel considerare materiali *omogenei e isotropi* e nel supporre la seguente relazione *lineare* tra i campi: $\vec{M} = \chi_m \vec{H}$. La costante scalare adimensionata χ_m (normalmente nota come *suscettività magnetica*) misura proprio la risposta "magnetica" del materiale al campo \vec{H} che vi è applicato. Unendo le relazioni di definizione scritte sopra, si trova facilmente che, nelle ipotesi fatte, si ha $\vec{B} = \mu_0 \mu_r \vec{H}$, con $\mu_r = 1 + \chi_m$ *permeabilità magnetica* del materiale.

A. Ferromagneti

Nella maggioranza dei materiali $\chi_m \approx 0$, con un segno positivo o negativo a seconda del carattere *para-* o *diamagnetico*. Di conseguenza gli effetti in termini magnetici della maggioranza dei materiali sono estremamente ridotti, benché i fenomeni che ne derivano siano estremamente importanti in molti settori della fisica della materia. In altre parole, la misura, ovvero la semplice verifica, delle proprietà magnetiche della materia per diamagneti e paramagneti non può essere condotta con i metodi tipici dei nostri esperimenti a causa, principalmente, della scarsa sensibilità e del ridotto rapporto segnale/rumore tipico delle misure elettriche macroscopiche. Completamente diverso è il caso dei materiali classificati come *ferromagnetici* (ad esempio, materiali ferrosi o contenenti Ni, Co, alcuni ioni di terre rare, e poco altro, di naturale o artificiale). Qui χ_m , e quindi μ_r , possono assumere valori dell'ordine delle migliaia e anche di parecchie decine di migliaia e quindi dare luogo a effetti ben visibili anche con semplici misure elettriche macroscopiche.

Nei ferromagneti l'eclatante valore di μ_r comporta un altrettanto eclatante aumento dell'intensità del campo \vec{B} rispetto al vuoto (o a materiali non ferromagnetici), oltre a diverse altre conseguenze tra cui il fenomeno di "canalizzazione" delle linee di \vec{B} all'interno del materiale che è stato già discusso in un'altra nota. Accanto a tutto questo, normalmente si ha anche un altro importante ef-

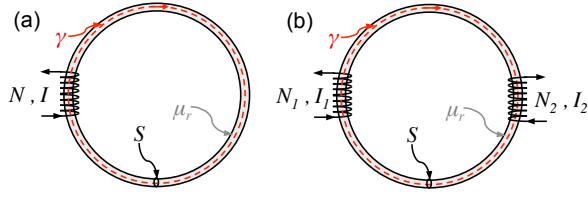


Figura 1. Circuito magnetico con uno (a) e due (b) avvolgimenti. La lunghezza della curva γ è ℓ , secondo quanto stabilito nel testo, dove l'area da essa racchiusa è indicata come Σ .

fetto: la relazione di linearità tra i campi che abbiano prima supposto non è più verificata, o, perlomeno, è verificata in termini di definizione solo assumendo che χ_m e μ_r siano una *funzione*, parecchio complicata e non univocamente determinata, dell'intensità H . Questo comportamento dà luogo alla formazione di un *ciclo di isteresi* nelle relazioni tra le ampiezze dei campi \vec{M} (o \vec{B}) e l'ampiezza di \vec{H} . In questa nota supporremo, salvo quando diversamente specificato, che la relazione tra i vari campi sia lineare, cioè che μ_r abbia un valore determinato univocamente per il materiale considerato (e nelle condizioni considerate).

II. CIRCUITI MAGNETICI

Supponiamo di avere un anello di materiale ferromagnetico: per semplicità, immaginiamo che il materiale sia omogeneo (μ_r uniforme dappertutto all'interno del materiale), che l'anello abbia una sezione uniforme S e che ℓ sia la sua lunghezza; dal punto di vista geometrico tale quantità non è ben definita a causa dello spessore dell'anello, che è non nullo se $S \neq 0$. Tuttavia possiamo sicuramente intendere con ℓ la lunghezza "media". Se preferite, potete pure assumere che $\ell >> \sqrt{S}$, cioè che l'anello abbia uno spessore così sottile che la lunghezza effettiva corrisponda al primo ordine con quella "media". Tenete comunque presente che tutte queste considerazioni e approssimazioni hanno il solo scopo di semplificare la trattazione matematica e che esse non incidono sulla generalità delle affermazioni riportate in questa sezione.

A causa dell'effetto di canalizzazione delle linee di campo di induzione magnetica, un eventuale campo \vec{B} che si trovi all'interno dell'anello rimane confinato in esso. In altre parole, si verifica che le linee del campo \vec{B} sono tutte contenute nell'anello: in queste condizioni, il flusso $\Phi_S(\vec{B})$ calcolato sulla sezione di anello è *invariante* per tutto l'anello, che quindi rappresenta un *tubo di flusso* richiuso su se stesso. Aggiungiamo ora alla nostra descrizione una bobina di N spire percorse da una corrente di intensità I e avvolte attorno all'anello di materiale ferromagnetico, come rappresentato in Fig. 1(a). La bobina produce un campo magnetico e quindi si comporta come generatore per le linee di campo \vec{B} che restano contenute nell'anello e si richiudono su se stesse.

La configurazione descritta assomiglia concettualmente a un tubo richiuso su se stesso e comprendente una pompa che fa circolare dell'acqua: anche in questo caso è agevole individuare un campo vettoriale, quello della velocità dei volumetti di fluido, il cui flusso sulla sezione del tubo è invariante. Un'analogia ancora più rilevante è quella di un circuito elettrico che, nella sua forma più semplice, può essere immaginato come un filo, dotato di una certa resistenza, che si chiude su un generatore di d.d.p.: qui a essere invariante su tutto il circuito è l'intensità di corrente (se volete, la carica elettrica "si conserva"), cioè il flusso $\Phi_S(\vec{J})$, con \vec{J} densità di corrente. A causa di questa analogia, all'anello di materiale ferromagnetico chiuso su se stesso si dà il nome di *circuito magnetico*.

Avere una configurazione sperimentale in cui si sa dove si trovano le linee di campo magnetico (cioè si sa che ci sono delle regioni in cui il campo di induzione magnetica è nullo) rappresenta una situazione pressoché impossibile da verificarsi in assenza di circuiti magnetici. Essa dà dei vantaggi fondamentali per il modello e il calcolo delle grandezze fisiche di interesse, in modo simile al concetto totalmente irrealistico e impossibile da realizzare del vituperato (da me) solenoide infinito. Usando lo strumento dei circuiti magnetici, che, per altro, esistono e funzionano nella realtà, non c'è alcun bisogno di inventarsi solenoidi infiniti.

Consideriamo dunque l'anello magnetico con la bobina avvolta attorno. Nella forma integrata sulla superficie, l'equazione di Maxwell del rotore di \vec{H} recita, sempre nel caso stazionario,

$$\oint_{\gamma} \vec{H} \cdot d\vec{\ell} = \Phi_{\Sigma}(\vec{J}), \quad (1)$$

dove γ rappresenta un perimetro della superficie Σ su cui si calcola il flusso della densità di corrente (tale flusso è anche noto come *corrente concatenata*). Supponendo valida la condizione di relazione isotropa tra \vec{H} e \vec{B} nel ferromagnete, le linee di campo di \vec{H} sono parallele a quelle di \vec{B} . Dunque esistono delle linee chiuse γ che si trovano all'interno del materiale ferromagnetico in cui \vec{H} e $d\vec{\ell}$ sono paralleli (eventualmente antiparalleli, a seconda del verso di circuitazione) fra loro, per cui l'Eq. 2 diventa

$$\oint_{\gamma} H d\ell = \Phi_{\Sigma}(\vec{J}) = NI, \quad (2)$$

dove nell'ultimo passaggio abbiamo anche usato la circostanza che la corrente concatenata, cioè quella che attraversa una superficie delimitata dalla curva γ , è pari a N -volte l'intensità di corrente I che passa nell'avvolgimento.

Per semplificare la matematica e togliere di torno i segni di integrale possiamo a questo punto utilizzare le approssimazioni geometriche riportate all'inizio di questa sezione, in particolare supporre che sia possibile individuare univocamente la lunghezza ℓ del circuito di integrazione e che la sezione dell'anello rimanga uniformemente

al valore S su tutta la sua lunghezza. Come ulteriore e ragionevole affermazione che discende dalle approssimazioni in uso, possiamo per semplicità supporre che il campo di induzione magnetica \vec{B} sia uniforme sull'intera sezione dell'anello, cioè che $\Phi_S(\vec{B}) = BS$, e anche lungo tutta la sua lunghezza, vista l'omogeneità del materiale. In queste condizioni si ha

$$\oint_{\gamma} H d\ell = H \ell = \frac{B \ell}{\mu_0 \mu_r} = \Phi_S(\vec{B}) \frac{\ell}{\mu_0 \mu_r S} = \Phi_S(\vec{B}) \mathcal{R} = NI . \quad (3)$$

L'Eq. 3 definisce la grandezza \mathcal{R} , opportunamente dimensionata, che dipende unicamente dalla geometria (forma e dimensioni) e dal materiale dell'anello. A questa grandezza si dà il nome di *riluttanza magnetica*. È ovvio che la sua semplice definizione, basata sulla dipendenza lineare diretta con la lunghezza del circuito magnetico e inversa con area della sezione e permeabilità magnetica, può diventare più complicata se qualcuna delle approssimazioni prima adottate viene rilassata, però concettualmente è sempre possibile individuare la riluttanza di un circuito magnetico e legarla appunto alle sue caratteristiche costruttive.

La riluttanza magnetica così definita ricorda molto da vicino la resistenza definita per un circuito elettrico. Nel caso di conduttore con resistività ρ_c uniforme e di simmetria "piana" (campo elettrico uniforme), la resistenza di un pezzo di filo di lunghezza ℓ e sezione S vale proprio $R = \rho_c \ell / S$, un'espressione che è formalmente analoga alla definizione di riluttanza in Eq. 3, in particolare nella dipendenza lineare diretta con la lunghezza e quella inversa con l'area della sezione. Dato che la definizione di resistenza è anche (impropriamente) citata come legge di Ohm, è possibile costruire una "legge" (il virgolettato serve a ricordarsi che in questi casi la definizione di legge è dovuta a processi storici con basi sperimentali, dato che le leggi di cui stiamo parlando nascono da altre leggi, per esempio le equazioni di Maxwell, unite e definizioni e semplici modelli) formalmente simile che vale per i circuiti magnetici. Questa legge prende il nome di *legge di Hopkinson* ed è qui sotto scritta in modo da permettere un confronto immediato con la legge di Ohm:

$$\text{Ohm: } IR = \Delta V \quad (4)$$

$$\text{Hopkinson: } \Phi_S(\vec{B}) \mathcal{R} = NI . \quad (5)$$

In questo contesto il flusso del campo di induzione magnetica nel circuito magnetico prende il ruolo dell'intensità di corrente (ovvero il flusso della densità di corrente) nel circuito elettrico, la riluttanza magnetica quello della resistenza elettrica e il prodotto NI quello della d.d.p.: per ribadire ancor meglio l'analogia, in qualche (vecchio) testo di fisica, dove si usa il termine forza elettromotrice per definire, nelle opportune condizioni, una d.d.p., il prodotto NI prende il nome un po' demodé (e misleading) di *forza magnetomotrice*.

A. Riluttanza e coefficienti di induzione magnetica

La possibilità di determinare i campi magnetici rilevanti usando circuiti magnetici si presta a stabilire un'importante relazione che coinvolge i coefficienti di auto e mutua induzione. Ricordiamo che essi sono definiti come:

$$L \equiv \frac{\Phi_{S,avv}(\vec{B})}{I} \quad (6)$$

$$M \equiv \frac{\Phi_{S,avv1}(\vec{B}_2)}{I_2} = \frac{\Phi_{S,avv2}(\vec{B}_1)}{I_1} , \quad (7)$$

dove tutte le grandezze e simbologie sono già state discusse altrove, a parte il dettaglio, altrove trascurato, che i flussi dei campi di induzione magnetica devono essere intesi come calcolati sull'"area dell'avvolgimento", da cui il pedice "avv".

Cominciamo con il considerare il caso di un singolo avvolgimento di N spire percorso da una corrente di intensità I realizzato su un circuito magnetico, come in Fig. 1(a). L'Eq. 5 stabilisce

$$\Phi_S(\vec{B}) = \frac{NI}{\mathcal{R}} , \quad (8)$$

dove \mathcal{R} è una certa (cioè nota) riluttanza del circuito magnetico. Come già stabilito, il flusso del campo di induzione magnetica nell'Eq. 8 si intende calcolato sulla sezione del circuito magnetico, che corrisponde alla sezione di *una singola spira* dell'avvolgimento. Ricordando la legge di Faraday, si ha che la "forza elettromotrice" indotta da un campo variabile nel tempo su una singola spira dell'avvolgimento (ovvero, senza entrare nella discussione sui segni, la d.d.p. ai capi della spira) è $\Delta V_{spira} = d\Phi_S(\vec{B})/dt$. Le spire sono tutte collegate in serie tra loro, per cui $\Delta V_{avv} = N\Delta V_{spira}$. Quindi $\Delta V_{avv} = N\Delta V_{spira} = Nd\Phi_S(\vec{B})/dt$, relazione che autorizza la conclusione, molto poco rigorosa in termini formali, $\Phi_{S,avv}(\vec{B}) = N\Phi_S(\vec{B})$.

Dunque, sfruttando l'Eq. 8 l'Eq. 6 si può scrivere:

$$L \equiv \frac{N\Phi_S(\vec{B})}{I} = \frac{N^2}{\mathcal{R}} , \quad (9)$$

che stabilisce per L una dipendenza *quadratica con il numero delle spire*. Osservate bene come questa dipendenza, benchè sia stata trovata supponendo la presenza di un circuito magnetico, vale anche per avvolgimenti realizzati nell'aria o su materiali non ferromagnetici. Infatti la definizione di riluttanza magnetica si basa sulla circuitazione di \vec{H} che può essere comunque realizzata anche in assenza di circuiti magnetici. L'unica, non marginale, differenza è che, se il campo magnetico non si trova confinato in un volume di geometria e materiale con permeabilità magnetica nota (e molto alta), \mathcal{R} non può essere calcolato. Anche in questo caso vedete come appellarsi a situazioni completamente irrealistiche, tipo solenoidi infiniti, in cui ovviamente si ottiene la stessa dipendenza, è del tutto inutile.

Passiamo ora alla mutua induzione e consideriamo il circuito magnetico di Fig. 1(b) dove sono presenti due avvolgimenti con numero di spire rispettivamente N_1 e N_2 percorsi da correnti di intensità I_1 e I_2 . L'invarianza di $\Phi_S(\vec{B})$ su tutto il circuito magnetico, unita alla considerazione appena fatta sui flussi calcolati sulle sezioni del circuito magnetico e sugli avvolgimenti, permette per esempio di porre: $\Phi_{S,avv1}(\vec{B}_2) = N_1\Phi_S(\vec{B}_2)$, per cui l'Eq. 7 si può scrivere:

$$M \equiv \frac{\Phi_{S,avv1}(\vec{B}_2)}{I_2} = N_1 \frac{\Phi_S(\vec{B}_2)}{I_2} = \frac{N_1 N_2}{\mathcal{R}}, \quad (10)$$

che stabilisce per M una dipendenza *lineare con il prodotto* del numero di spire dei due avvolgimenti; valgono ovviamente tutte le considerazioni appena svolte per l'autoinduttanza, compresa l'inutilità di pensare a irrealistiche configurazioni di solenoidi infiniti coassiali, e anche di configurazioni toroidali coassiali (meno irrealistiche, ma ugualmente poco utili).

B. Applicazioni di circuiti magnetici e Hopkinson

Ci sono alcune applicazioni rilevanti dei concetti fin qui esposti che vale la pena di trattare brevemente.

La prima si riferisce alla realizzazione di *schermi magnetici* per campi stazionari. Già sappiamo che la schermatura di campi di induzione magnetica variabili nel tempo può efficacemente essere realizzata con gusci di materiale conduttore grazie all'azione delle *correnti parassite*, ma l'assenza degli effetti di induzione magnetica nel caso stazionario rende inutili tali schermi. Invece il fenomeno della canalizzazione delle linee di campo all'interno dei materiali ferromagnetici apre la possibilità di ridurre l'ampiezza dei campi di induzione magnetica statici in determinate regioni di spazio. L'effetto è schematizzato in Fig. 2(a), dove si suppone che un guscio cilindrico spesso di materiale con $\mu_r >> 1$, il cui asse è ortogonale alla figura, venga immerso in una regione di spazio in cui insiste un campo di induzione magnetica \vec{B} . Le linee del campo, che incidono con un certo angolo sulla superficie esterna del guscio, all'interno del materiale vengono "piegate" in modo da disporsi quasi parallelamente alla superficie di interfaccia. In questo modo esse vengono escluse dal volume interno al guscio, per riemergere al di fuori con un meccanismo che assomiglia a quello della rifrazione dei raggi luminosi che incidono su interfacce tra materiali dielettrici diversi.

Ovviamente la schermatura è tanto più efficace quanto maggiore è la permeabilità magnetica. Inoltre, per la realizzazione pratica degli schermi, occorre anche che il materiale sia lavorabile, una richiesta che è particolarmente rilevante vista la "delicatezza" delle leghe ferromagnetiche: infatti μ_r può diminuire sensibilmente se la temperatura aumenta durante la lavorazione o se essa implica l'applicazione di stress, o strain, particolarmente elevati. Tra i materiali più frequentemente usati c'è il

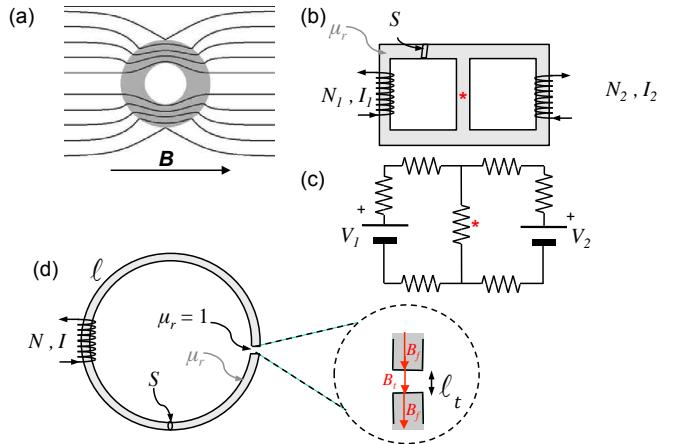


Figura 2. Illustrazione di alcune conseguenze e applicazioni rilevanti per circuiti magnetici e legge di Hopkinson secondo quanto discusso nel testo. I disegni non sono in scala.

mu-metal (si intuiscono facilmente le origini del nome), una lega con componenti principali Ni e Fe che garantisce $\mu_r > 10^3$ e fino ad alcune unità in 10^4 , caratteristica unita a una buona lavorabilità.

Un'altra applicazione rilevante ha a che fare con la forma della legge di Hopkinson. Poiché essa è analoga a quella della legge di Ohm, ci si può aspettare che tutte le "tecniche" usate nella soluzione dei circuiti elettrici, basate o compatibili con la legge di Ohm, possano essere trasferite ai circuiti magnetici. Tra queste "tecniche" ci sono sicuramente la descrizione dei circuiti in termini di nodi, rami, maglie, e le regoline che si applicano a nodi (somma delle intensità di corrente nulla) e a maglie (somma delle d.d.p. nulla), chiamate pomposamente "leggi" di Kirchoff. È infatti possibile ipotizzare circuiti magnetici "complicati", come per esempio quello in Fig. 2(b), di cui supponiamo di conoscere le caratteristiche geometriche (lunghezze e sezione dei vari pezzi) e di permeabilità magnetica. L'analogico in termini di circuiti elettrici è disegnato schematicamente in Fig. 2(c): i tratti di materiale ferromagnetico, dotati da una certa riluttanza, corrispondono ai rami resistivi del circuito elettrico, esistono dei nodi in cui il flusso del campo di induzione magnetica si dirama, esattamente come si dirama l'intensità di corrente ai due nodi del circuito elettrico, e infine i due avvolgimenti, di N_1 e N_2 spire percorse da corrente di intensità I_1 e I_2 , corrispondono a due distinti generatori (ideali) di d.d.p., V_1 e V_2 . Nel caso elettrico sappiamo come procedere per risolvere il circuito, cioè ottenere le informazioni necessarie a determinare le grandezze rilevanti (d.d.p. e intensità di corrente) in tutto il circuito: è sufficiente individuare (due) maglie indipendenti e scrivere le rispettive equazioni lineari che legano d.d.p. alle intensità di corrente tramite il valore delle resistenze, usando, quando necessario, il principio di sovrapposizione. Lo stesso approccio può essere impiegato nel caso del circuito magnetico: potete provare per esercizio a determinare il flusso di campo magnetico presente nel ramo

“centrale” del circuito, quello marcato con un asterisco, che corrisponde formalmente a determinare l’intensità di corrente che fluisce nel ramo con la resistenza asteriscata del circuito elettrico “equivalente”.

Infine, un’ulteriore applicazione fa riferimento ai cosiddetti *traferri*, regioni di spazio vuoto che si ottengono rimuovendo una piccola porzione di un circuito magnetico come indicato in Fig. 2(d). Per semplicità, immaginiamo che la rimozione avvenga in modo da lasciare esposte due interfacce parallele tra materiale ferromagnetico e aria, poste a distanza relativa $\ell_t \ll \ell$, con ℓ lunghezza complessiva del circuito magnetico prima della formazione del traferro. Poichè la geometria del sistema, tenendo anche conto delle assunzioni di Sez. II, implica che lo spessore del traferro sia molto minore della dimensione trasversale del materiale ferromagnetico che realizza il circuito, si possono ragionevolmente trascurare gli “effetti ai bordi”, per cui la direzione del campo di induzione magnetica è sempre ortogonale rispetto alle interfacce. Di conseguenza l’ampiezza del campo nel traferro, B_t , è per continuità pari a quella nel ferro, B_f ; essendo pari anche le aree delle sezioni di interesse, si ha anche $\Phi_S(\vec{B}_f) = \Phi_S(\vec{B}_t)$, circostanza che permette di considerare in serie tra di loro le riluttanze del ferro, \mathcal{R}_f , e del traferro, \mathcal{R}_t . Tenendo conto di geometria e materiali, si ha

$$\mathcal{R}_f = \frac{\ell - \ell_t}{\mu_0 \mu_r S} \quad (11)$$

$$\mathcal{R}_t = \frac{\ell_t}{\mu_0 S} \quad (12)$$

$$\begin{aligned} \mathcal{R}_{eq} &= \mathcal{R}_f + \mathcal{R}_t = \frac{\ell}{\mu_0 \mu_r S} \left(1 + \frac{\ell_t(\mu_r - 1)}{\ell} \right) \simeq \\ &\simeq \frac{\ell}{\mu_0 \mu_r S} \left(1 + \mu_r \frac{\ell_t}{\ell} \right), \end{aligned} \quad (13)$$

dove abbiamo espresso come riluttanza equivalente la somma delle due riluttanze e fatto un’approssimazione al primo ordine lecita se $\mu_r \gg 1$, come stiamo supponendo.

Per la legge di Hopkinson si ha quindi che il campo di induzione magnetica nel traferro ha ampiezza approssimata

$$B_t = \frac{NI}{\mathcal{R}_{eq} S} \simeq NI \frac{\mu_0 \mu_r}{\ell + \mu_r \ell_t}, \quad (14)$$

che mostra come, per almeno per traferri sottili, il campo di induzione magnetica nel traferro, cioè nel vuoto, possa assumere un’ampiezza rilevante sfruttando “per continuità” l’“amplificazione” dovuta al materiale ferromagnetico.

III. TRASFORMATORE

L’utilità pratica dei dispositivi che vanno sotto il nome di trasformatori è più che evidente e nota a tutti. Molti apparecchi elettrici, e praticamente tutti i dispositivi

elettronici di cui facciamo continuamente uso, hanno bisogno di un’alimentazione in corrente continua e a bassa tensione.

Una tensione, o corrente, di questo tipo non è però quella che esce dalle normali prese di casa, cioè dalla rete di distribuzione dell’energia elettrica. In primo luogo, la rete elettrica fornisce corrente alternata. Il motivo è soprattutto perché di questo tipo è la differenza di potenziale prodotta dai generatori (alternatori) che tipicamente sfruttano un moto periodico e, in accordo con la legge di Faraday, forniscono una forza elettromotrice alternata. Inoltre l’ampiezza di questa forza elettromotrice, che ha forma approssimativamente sinusoidale e frequenza $f = 50$ Hz, è normalmente elevata ($230 - 240$ V_{rms} per la rete di casa, fino a centinaia di kV per gli elettrodotti ad alta tensione). Il motivo è legato al desiderio di minimizzare le perdite per effetto Joule da parte dei fili che trasportano la corrente in giro per il mondo. Infatti, data una certa resistenza (senz’altro non trascurabile se si considera la lunghezza dei fili necessari per cablare il territorio), la potenza dissipata scala con il quadrato della corrente. A parità di potenza fornita dal generatore (dalla centrale elettrica), la corrente scala inversamente con la tensione. Dunque le perdite per effetto Joule diminuiscono con il quadrato della tensione, permettendo di limitare lo spreco di energia.

Conoscete già un modo per convertire in continua una tensione alternata, costituito dall’unione del raddrizzatore a diodo e del livellatore con condensatore (ovvero integratore RC). Per abbassare l’ampiezza della d.d.p. alternata che deve poi essere sottoposta a raddrizzamento e livellamento ci si serve prevalentemente proprio dei trasformatori, che per questo hanno un impiego diffusissimo. Vale tuttavia la pena di ricordare che, grazie allo sviluppo dell’elettronica, sono via via entrati in uso dei dispositivi che svolgono la stessa funzione facendo uso di modalità di operazione leggermente differenti. Questi dispositivi, che si definiscono *switching* e che costituiscono la quasi totalità degli attuali alimentatori per piccoli apparecchi elettronici (telefonini, computer, etc.), funzionano attraverso livellamento di un’onda quadra che opera a frequenze nettamente maggiori rispetto ai tradizionali 50 Hz (tipicamente decine di kHz) e che ha un *duty cycle* variabile in funzione della richiesta di corrente da parte dell’apparato che deve essere alimentato. In questo modo, come in parte vedremo nel seguito di questa nota, essi permettono di aumentare il rendimento e minimizzare le perdite di potenza. In ogni caso anche al loro interno si trovano dei componenti che assomigliano molto da vicino al trasformatore convenzionale, sul quale concentriamo la nostra attenzione.

Un trasformatore è, in un’ampia accezione, un sistema costituito da due avvolgimenti (*primario e secondario*) *avvolti sullo stesso circuito magnetico*. La Fig. 3 mostra lo schema di qualche realizzazione costruttiva. L’uso di circuiti magnetici è l’unica possibilità pratica realistica che consente di avere un accoppiamento pressoché completo tra i due avvolgimenti, cioè realizzare un *coefficiente*

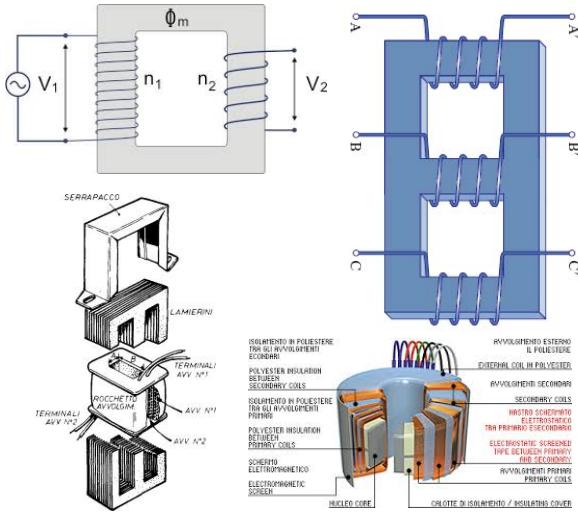


Figura 3. Schemini realizzativi di trasformatori trovati in rete. In ordine orario, dall'alto a sinistra: trasformatore a due avvolgimenti, trasformatore a tre avvolgimenti (per esempio per impieghi con d.d.p. “trifase”), trasformatore toroidale a uscite multiple, esplosione di trasformatore a due avvolgimenti.

di accoppiamento magnetico $k \simeq 1$ (per esempio $k \gtrsim 0.95$ per avvolgimenti toroidali su anelli di acciai magnetici, in genere leghe ferrose contenenti silicio). Un *trasformatore ideale*, che è piuttosto bene approssimato da tali realizzazioni pratiche, è quello per cui $k = 1$. Ricordando la definizione già data altrove, $k = 1$ implica

$$M = \sqrt{L_1 L_2} . \quad (15)$$

Nel trasformatore ideale che consideriamo per i nostri scopi supponiamo che il circuito magnetico sia realizzato con materiale di permeabilità μ_r e che i due avvolgimenti siano realizzati su sezioni di pari area S del circuito magnetico, del quale indichiamo con ℓ la lunghezza: in sostanza, il sistema ha la geometria di Fig. 1(b), anche se la forma del circuito potrebbe non essere circolare (per esempio quadrata o rettangolare, come in alcuni esempi di Fig. 3). Chiamiamo poi N_1 e N_2 il numero delle spire di secondario e primario e, per praticità, individuiamo il *parametro costruttivo*

$$\alpha = \frac{N_2}{N_1} . \quad (16)$$

Poiché il circuito magnetico ha un'unica riluttanza “vista” dai due avvolgimenti, $\mathcal{R} = \ell / (\mu_0 \mu_r S)$, si ha $L_i = \mu_0 \mu_r S N_i^2 / \ell$, con $i = 1, 2$; questa espressione è pericolosamente simile alla vituperata analoga formula che si ottiene con l’irrealistica configurazione del solenoide infinito (dove, per altro, $\ell \rightarrow \infty$), con la *non trascurabile* differenza che ℓ rappresenta qui la lunghezza del circuito magnetico e non del solenoide. Le Eqs. 9,10 combinate

alle Eqs. 15,16 stabiliscono

$$\frac{L_2}{L_1} = \left(\frac{N_2}{N_1} \right)^2 = \alpha^2 \quad (17)$$

$$\frac{M}{L_1} = \frac{N_2}{N_1} = \alpha \quad (18)$$

$$\frac{M}{L_2} = \frac{N_1}{N_2} = \frac{1}{\alpha} . \quad (19)$$

In altre parole, in un trasformatore le induttanze di primario e secondario sono legate tra di loro e con la mutua induzione attraverso semplici relazioni *costruttive*; inoltre, poiché normalmente il numero di spire è ben diverso tra primario e secondario, questo vuol dire che le induttanze sono anche molto differenti tra loro.

A. Trasformazione in tensione e in corrente

La configurazione sperimentale di riferimento è schematizzata in Fig. 4: i due avvolgimenti primario e secondario sono realizzati sullo stesso circuito magnetico, da cui la rappresentazione simbolica dello schema e nei rispettivi circuiti supponiamo di avere resistenze R_1 e R_2 , eventualmente comprensive delle resistenze interne degli induttori. Nell’impiego ordinario un generatore di d.d.p. alternata e periodica è collegato al primario a fornire V_1 , mentre un carico, eventualmente costituito da R_2 , è presente al secondario secondo quanto illustrato nel seguito. Le equazioni generiche dei circuiti primario e secondario per segnali sinusoidali a frequenza angolare ω possono essere scritte nel dominio delle frequenze come

$$V_{\omega,1} = (R_1 + j\omega L_1) I_{\omega,1} + j\omega M I_{\omega,2} \quad (20)$$

$$V_{\omega,2} = (R_2 + j\omega L_2) I_{\omega,2} + j\omega M I_{\omega,1} . \quad (21)$$

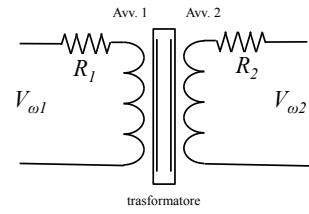


Figura 4. Rappresentazione schematica del circuito con trasformatore considerato nel testo.

Determiniamo ora il rapporto di trasformazione in tensione $T_V = V_2/V_1$, dove V_1 e V_2 sono le ampiezze, o ampiezze picco-picco, delle d.d.p. al primario e secondario. Come noto, a questo scopo possiamo operare nelle condizioni di *secondario aperto*, cioè porre $I_{\omega,2} = 0$, ottenendo con pochi passaggi documentati altrove, in cui si suppone $\omega L_1 \gg R_1$,

$$T_V = \frac{V_2}{V_1} = \frac{M}{L_1} = \frac{N_2}{N_1} = \alpha . \quad (22)$$

Nelle approssimazioni fatte si ha che T_V non dipende da ω (nell'ambito delle approssimazioni applicate) e inoltre, come si può facilmente verificare, che $V_{\omega,1}$ e $I_{\omega,1}$ sono sfasati tra di loro di $\pi/2$ rad, per cui *a secondario aperto il trasformatore non assorbe alcuna potenza* al di fuori di quella spesa per effetto Joule su R_1 . Nell'impiego ordinario di un trasformatore casalingo, in cui esso serve per *ridurre la tensione della rete elettrica collegata al primario*, N_1 è grande per cui alta è la sua induttanza L_1 . Di conseguenza, l'approssimazione sopra citata è facilmente verificata anche per frequenze relativamente basse.

Se vogliamo determinare il rapporto di trasformazione in corrente possiamo invece operare nelle condizioni di *secondario cortocircuitato* sul carico resistivo R_2 , cioè porre $V_{\omega,2} = 0$. Come discusso altrove, nell'ipotesi $\omega L_2 \gg R_2$ si ottiene che il rapporto di trasformazione in corrente $T_A = I_2/I_1$, con I_1 e I_2 ampiezze, o ampiezze picco-picco, delle intensità di corrente al primario e al secondario, si scrive

$$T_A = \frac{I_2}{I_1} = \frac{M}{L_2} = \frac{N_1}{N_2} = \frac{1}{\alpha}, \quad (23)$$

che è dunque il reciproco di T_V . Pertanto anche il rapporto di trasformazione in corrente è indipendente dalla frequenza (nell'ambito delle approssimazioni applicate) e si può facilmente verificare che le correnti di intensità $I_{\omega,1}$ e $I_{\omega,2}$ sono sfasate tra di loro di $\pm\pi$ rad, dove l'incertezza sul segno dipende anche dalla circostanza che in genere non si sa se gli avvolgimenti sono realizzati in senso concorde o discorde tra loro.

L'operazione a secondario cortocircuitato comporta anche, come già stabilito in una precedente nota,

$$\begin{aligned} V_{\omega,1} &= \left(R_1 + j\omega L_1 + \frac{\omega^2 M^2}{R_2 + j\omega L_2} \right) I_{\omega,1} = \\ &= \left(R_1 + \frac{-\omega^2 L_1 L_2 + \omega^2 M^2 + j\omega L_1 R_2}{R_2 + j\omega L_2} \right) I_{\omega,1} = \\ &= \left(R_1 + \frac{j\omega L_1 R_2}{R_2 + j\omega L_2} \right) I_{\omega,1}, \end{aligned} \quad (24)$$

dove abbiamo sfruttato l'ipotesi di Eq. 15. Nella solita approssimazione $\omega L_2 \gg R_2$ si ottiene allora:

$$\begin{aligned} V_{\omega,1} &= \left(R_1 + \frac{L_1}{L_2} R_2 \right) I_{\omega,1} = \left(R_1 + \frac{N_1^2}{N_2^2} R_2 \right) I_{\omega,1} = \\ &= \left(R_1 + \frac{1}{\alpha^2} R_2 \right) I_{\omega,1}. \end{aligned} \quad (25)$$

Questa equazione è molto importante poiché ci dice che, nelle approssimazioni fatte, il primario si comporta in modo esclusivamente resistivo, cioè la corrente è in fase con la d.d.p., e come se fosse dotato di una resistenza efficace $R_{eff} = R_1 + R_2/\alpha^2$.

B. Rendimento del trasformatore ideale

Il rendimento η di un trasformatore ha a che fare con la capacità del dispositivo di trasferire potenza dal pri-

mario, dove normalmente è collegato un generatore di d.d.p. in grado di erogare potenza, al secondario, dove supponiamo sia presente il carico resistivo R_2 . Allo scopo di definire una grandezza capace di stabilire l'efficienza della configurazione di avvolgimenti senza contaminazioni dovute alla dissipazione resistiva al primario, il rendimento η è definito come

$$\eta = \frac{P_2}{P_1 - P_{J1}}, \quad (26)$$

dove tutte le potenze espresse vanno intese come *mediate nel tempo*, P_2 e P_1 rappresentano le potenze al primario e al secondario, P_{J1} è la potenza dissipata per effetto Joule sulla resistenza R_1 del primario.

Le potenze P_{J1} e P_2 sono utilizzate da carichi resistivi, per cui esse possono essere espresse come

$$P_{J1} = \frac{R_1 I_1^2}{2} \quad (27)$$

$$P_2 = \frac{R_2 I_2^2}{2} = \frac{R_2 I_1^2}{2\alpha^2}, \quad (28)$$

dove abbiamo sfruttato il rapporto di trasformazione in corrente espresso in Eq. 23.

In termini generali la potenza P_1 al primario, che è quella (segni a parte) erogata dal generatore, si scrive

$$P_1 = \frac{1}{2} \operatorname{Re}\{V_{\omega,1} \cdot I_{\omega,1}^*\} = \frac{V_1 I_1}{2} \cos(\Delta\phi), \quad (29)$$

dove $\Delta\phi$ è lo sfasamento tra $V_{\omega,1}$ e $I_{\omega,1}$. Nelle approssimazioni utilizzate, l'Eq. 25 stabilisce che $\Delta\phi = 0$ e che il carico visto dal primario, dunque dal generatore ad esso collegato, equivale alla resistenza efficace $R_{eff} = R_1 + R_2/\alpha^2$. Pertanto è $V_1 = R_{eff} I_1$ e si ha

$$P_1 = \frac{(R_1 + R_2/\alpha^2) I_1^2}{2}, \quad (30)$$

essendo $\cos(\Delta\phi) = 1$.

Mettendo le Eqs. 27,28,30 nella definizione di Eq. 26 si ottiene

$$\begin{aligned} \eta &= \frac{\frac{R_2 I_1^2}{2\alpha^2}}{\frac{R_1 I_1^2}{2} + \frac{R_2 I_1^2}{2\alpha^2} - \frac{R_1 I_1^2}{2}} = \\ &= \frac{\frac{R_2 I_1^2}{2\alpha^2}}{\frac{R_2 I_1^2}{2\alpha^2}} = 1, \end{aligned} \quad (31)$$

che dimostra come il trasformatore ideale abbia *rendimento unitario*. Osservate che questa conclusione è in accordo con la relazione tra rapporti di trasformazione in tensione e corrente, $T_V T_A = 1$, che abbiamo verificato in Sez. III A. Infatti, se, ad esempio, al secondario la d.d.p. si riduce rispetto al primario, la corrente disponibile aumenta, ma il prodotto tra tensione e intensità di corrente al primario e al secondario, che rappresenta le potenze coinvolte, deve rimanere invariato.

La potenza del generatore è ovviamente tanto meglio accoppiata al carico che si trova al secondario quanto minori sono le perdite per effetto Joule al primario. Per verificare l'efficacia dell'accoppiamento, riscriviamo l'Eq. 28 in funzione di V_1 :

$$P_2 = \frac{V_1^2}{2} \frac{R_2/\alpha^2}{(R_1 + R_2/\alpha^2)^2}, \quad (32)$$

dove abbiamo usato la $I_1 = V_1/R_{eff}$. Ponendo V_1 costante, come si verifica se il generatore di d.d.p. alternata è considerato *ideale*, quella appena scritta può essere considerata come una funzione di R_2 , che va a zero per $R_2 \rightarrow 0$ e $R_2 \rightarrow \infty$, e ha un *massimo* per $R_2 = \alpha^2 R_1$ (tutto questo può essere facilmente verificato con pochi passaggi matematici). Dunque il trasferimento di potenza dal primario al secondario è massimizzato per un determinato valore di R_2 , in corrispondenza del quale, a parità di tutti gli altri parametri, la potenza persa per effetto Joule nel circuito del primario è minimizzata. È interessante notare che in queste condizioni si ha $R_{eff} = 2R_1$, cioè la resistenza efficace del primario è pari al doppio di R_1 . Di conseguenza, e sempre supponendo condizioni ideali, metà della potenza fornita dal generatore viene spesa per effetti di natura resistiva (finisce in P_{J1}) e metà per effetti di natura reattiva, cioè legati a auto e mutua induzione. Si verifica, quindi, una condizione di matching tra le resistenze della sorgente e del carico, simile, almeno dal punto di vista concettuale, a quella di un generatore di Thévenin collegato a un carico resistivo.

C. Rendimento in trasformatori reali

Il rendimento unitario del trasformatore ideale dipende dalla validità delle approssimazioni e assunzioni che sono state applicate. Si può facilmente dimostrare che, se esse vengono rilassate, il rendimento diventa sub-unitario. Vale la pena di riassumerle:

1. l'accoppiamento magnetico è stato supposto unitario, cioè $k = 1$;
2. dove necessario, le componenti resistive delle impedenze al primario e al secondario sono state trascurate rispetto alle componenti reattive;
3. è stata trascurata qualsiasi altra possibile dissipazione o perdita di energia.

La condizione 1 dipende dalla qualità costruttiva dell'insieme di avvolgimenti e circuito magnetico. Come già affermato, è possibile realizzare trasformatori con $k \gtrsim 0.95$ usando geometrie toroidali e materiali ferromagnetici con alta permeabilità magnetica. Le configurazioni ottenute minimizzano il numero di linee di campo magnetico che, prodotte dal primario, non si concatenano con il secondario.

La condizione 2 dipende ovviamente dalla frequenza di operazione e dal carico resistivo collegato al secondario.

Nei trasformatori convenzionali, dove il numero di spire al secondario è relativamente piccolo e quindi piccola è l'induttanza, operando a $f = 50$ Hz l'approssimazione $\omega L_2 \gg R_2$ è in genere poco verificata. Uno dei vantaggi principali degli alimentatori switching è proprio nell'aumento della frequenza di operazione (fino a decine di kHz), che rende ragionevole l'approssimazione.

Per quanto riguarda la condizione 3, sappiamo bene che esistono diverse ulteriori meccanismi dissipativi coinvolti nel funzionamento del trasformatore. I principali sono legati all'isteresi ferromagnetica e alla presenza di correnti parassite nel circuito magnetico (è ragionevole trascurare la dissipazione associata all'effetto di prossimità citato in una precedente nota, oppure all'effetto pelle).

L'isteresi è inevitabilmente coinvolta nell'operazione del trasformatore, poiché la d.d.p. alternata in ingresso fa compiere periodicamente dei cicli di magnetizzazione/smagnetizzazione al materiale del circuito magnetico. Per limitare l'energia persa occorre che la curva di isteresi sia stretta, cioè che il materiale appartenga alla categoria convenzionale riferita spesso ai "ferri dolci". Gli acciai speciali contenenti silicio generalmente usati nei trasformatori soddisfano abbastanza bene questo requisito; tuttavia l'energia dissipata per isteresi dipende dal volume di materiale ferromagnetico coinvolto, che, nel caso di trasformatori ad alta potenza (di grandi dimensioni), è rilevante. In queste condizioni si può verificare che una frazione non trascurabile della potenza massima (fino a qualche parte per cento) che il trasformatore è in grado di gestire prima che si verifichino problemi di surriscaldamento sia spesa per percorrere il ciclo di isteresi.

Infine, l'effetto delle correnti parassite è quello che potenzialmente può dare luogo alla massima dissipazione di potenza spuria, producendo effetti certamente non trascurabili sul rendimento. Come abbiamo verificato in una precedente occasione, per dissipare potenza le correnti parassite devono circolare su linee chiuse sufficientemente estese, cioè tali da racchiudere superfici macroscopiche. Infatti la dissipazione creata da correnti parassite che si sviluppano alla superficie di un blocco pieno di materiale conduttore è molto più severa che non quella dovuta alla presenza di correnti parassite su lamine sottili. Praticamente tutti i circuiti magnetici impiegati nei trasformatori sono realizzati accoppiando lamine sottili *isolate* elettricamente tra di loro, in modo che le linee di corrente circoscrivano superfici relativamente piccole. Poiché le correnti parassite, come ogni corrente indotta, aumentano di intensità all'aumentare della frequenza, anche questo approccio può essere insufficiente quando la frequenza di operazione è elevata, per esempio negli alimentatori switching. In questi casi si usano frequentemente dei circuiti magnetici realizzati con materiali ad alta permeabilità e bassa conducibilità. Tra questi materiali i più rilevanti tecnologicamente sono le cosiddette *ferriti*, costituite nella pratica da dispersioni di ossidi di ferro (isolanti, o debolmente conduttori) in resine dielettriche.

Ponti

francesco.fuso@unipi.it; <http://www.df.unipi.it/~fuso/dida>

(Dated: version 3 - FF, 22 aprile 2016)

Questa nota tratta dell'argomento generale dei ponti di misura. L'argomento è molto interessante dal punto di vista storico e concettuale, benché non molto attuale e, purtroppo, non facilmente verificabile dal punto di vista sperimentale. Tuttavia è molto utile farne una breve trattazione.

I. INTRODUZIONE

L'uso delle configurazioni a ponte per misurare grandezze elettriche (resistenze, capacità, induttanze) ha un ruolo storico di rilievo. Infatti, come vedremo, la configurazione a ponte permette di aumentare la sensibilità della misura senza fare uso di componenti "attivi" (amplificatori di segnale), un po' come tutti quei sistemi che permettono di "esaltare" la precisione (pensate a uno strumento a lancetta con una lancetta molto lunga: un piccolo spostamento angolare diventa un grande spostamento tangenziale, facile da misurare).

Con lo sviluppo della tecnologia elettronica, analogica e digitale, i ponti di misura in quanto tali sono stati abbondantemente confinati a un ruolo "storico", dato che si sono resi disponibili metodi più raffinati per garantire la stessa sensibilità, o per averla addirittura migliore, evitando allo stesso tempo le complicazioni e la delicatezza strumentale che è intrinseca alle configurazioni a ponte. Per altro, complicazioni e delicatezza sono anche i motivi che rendono difficile realizzare praticamente esercitazioni sui ponti che abbiano un esito soddisfacente.

In ogni caso ci sono motivi concettuali e anche pratici (oltre che per la misura di grandezze elettriche, le configurazioni a ponte sono spesso usate per misurare il "bilanciamento" tra segnali, cioè per misurare piccole differenze tra segnali di bassa intensità) per cui è utile soffermarsi anche su questo argomento.

II. IL PONTE, IN GENERALE

In termini molto generali, la configurazione a ponte prevede di collegare (almeno) quattro componenti dotati di quattro impedenze, generalmente complesse, Z_{1-4} , secondo lo schema di Fig. 1. Il circuito prevede poi il collegamento a un generatore V_{in} , generalmente alternato e sinusoidale (quindi il problema è trattabile con il metodo simbolico), e la lettura di un segnale V_ω , anch'esso generalmente alternato e sinusoidale, secondo quanto riportato nello schema.

I rami A e B, costituiti rispettivamente dalle serie Z_1, Z_4 e Z_2, Z_3 , sono collegati tra loro in parallelo. Il segnale V_ω , è pari alla "caduta di potenziale" ai capi dell'impedenza 3 meno quella ai capi dell'impedenza 4, cioè: $V_\omega = Z_3 I_{\omega B} - Z_4 I_{\omega A}$, dove $I_{\omega A}$ e $I_{\omega B}$ sono i fasori delle intensità di corrente dei due rami. A causa del collegamento in parallelo tra di loro, si ha un partitore di corrente,

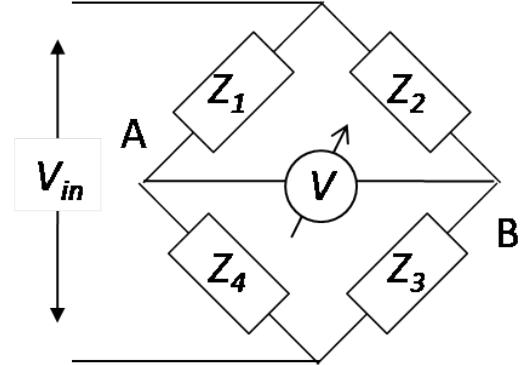


Figura 1. Topologia generale di una configurazione circuitale a ponte.

cioè $I_{\omega B}/I_{\omega A} = (Z_1 + Z_4)/(Z_2 + Z_3)$. Di conseguenza:

$$V_\omega = I_{\omega A} \left(Z_3 \frac{Z_1 + Z_4}{Z_2 + Z_3} - Z_4 \right) = \frac{I_{\omega A}}{Z_2 - Z_3} (Z_1 Z_3 - Z_2 Z_4). \quad (1)$$

Nell'uso della configurazione a ponte per eseguire misure si è normalmente interessati a determinare le condizioni in cui $|V_\omega| = 0$. Questo si verifica quando $Z_1 Z_3 = Z_2 Z_4$. Se guardate lo schema, vedete che questa equazione equivale a dire, a parole, che *sono uguali i prodotti delle impedenze che si trovano su lati "opposti" del ponte*. Per il momento, non ci preoccupiamo di specificare il significato matematico dell'equazione appena scritta nel caso in cui le impedenze considerate siano complesse, cioè non puramente reali o immaginarie. Vedremo caso per caso quali conseguenze si possono trarre dall'uguaglianza appena scritta.

Nel seguito esamineremo tre configurazioni a ponte particolarmente famose, che si differenziano fra loro per la tipologia di componenti impiegati e per le finalità della misura che è resa possibile.

A. Ponte di Wheatstone

Il ponte di Wheatstone è nient'altro che un ponte in cui le 4 impedenze sono tutte reali, cioè i 4 componenti sono 4 resistori (o composizioni di resistori). Visto che per i resistori non ci sono sfasamenti tra tensione e corrente, usare un generatore alternato non porta vantaggi rispetto a un generatore continuo, che è quello che immaginiamo di impiegare come V_{in} .

Supponiamo anche, come in genere si fa(ceva), che uno dei resistori abbia una resistenza incognita, per esempio R_1 , e che un altro dei resistori abbia un valore variabile, per esempio R_4 , che dunque potrebbe essere un *potenziometro*. La condizione di bilanciamento tra i rami, cioè quella per cui $|V_\omega| = 0$, si ottiene quando $R_1 = R_4 R_2 / R_3$. Quindi il valore, supposto incognito, di R_1 si può ottenere conoscendo quello degli altri resistori. In particolare, esso risulta dal valore di R_4 , che per esempio può essere aggiustato girando la manopola del potenziometro (che si può opportunamente calibrare misurando una volta per tutte la sua resistenza in funzione della rotazione della manopola), moltiplicato per il *fattore di scala* R_2/R_3 , che negli strumenti di misura poteva essere facilmente modificato per permettere di cambiare la portata. Scegliendolo in modo opportuno, la sensibilità di misura veniva opportunamente aumentata. Inoltre, grazie all'uso di strumenti (a lancetta, ovviamente) che potevano apprezzare differenze di potenziale, o correnti, di un segno o dell'altro, era sperimentalmente agevole determinare la condizione di bilanciamento osservando che la lancetta passasse proprio per la posizione di zero, tipicamente quella centrale.

Questo, a grandi linee, è quanto si faceva per e con un ponte di misura per le resistenze.

B. Ponte di de Sauty

In questo caso l'obiettivo è misurare una capacità. Immaginiamo dunque che un condensatore, di capacità incognita, occupi il posto del componente 1, la cui impedenza sarà quindi $Z_1 = 1/(j\omega C_1)$. È evidente che in questo caso si deve usare un generatore alternato. Infatti in continua un condensatore si comporta come un circuito aperto e non potrebbe passare corrente per il ramo A. La condizione di bilanciamento sarebbe possibile solo sostituendo con un circuito aperto almeno uno dei componenti del ramo B. Si avrebbe bilanciamento, ma, chiaramente, non si potrebbe determinare il valore di C_1 . Inoltre Z_1 è immaginaria, e quindi è necessario che almeno un altro componente del ponte abbia impedenza immaginaria, altrimenti il bilanciamento sarebbe impossibile da ottenere. Nella configurazione di de Sauty tale componente si trova nel ramo B, e noi supponiamo sia il componente 2, che è un condensatore di valore noto e ha quindi impedenza $Z_2 = 1/(j\omega C_2)$. Per gli altri due componenti sceglieremo ancora dei resistori, uno dei quali, per esempio il componente 4, supponiamo sia un potenziometro. La condizione di bilanciamento, da verificarsi ovviamente con uno strumento in grado di leggere segnali alternati, è allora $C_1 = C_2 R_3 / R_4$, che, agendo sul potenziometro R_4 , permette di ricavare C_1 a partire dalla conoscenza degli altri tre valori.

Notate un paio di aspetti molto importanti: si vede come la condizione di bilanciamento sia indipendente dalla frequenza. Naturalmente ciò si verifica solo se le capacità considerate sono indipendenti dalla frequenza, affer-

mazione che non è sempre e necessariamente del tutto vera. Dunque una misura con un ponte in cui si può modificare la frequenza del generatore costituisce un valido test finalizzato proprio a verificare entro quali limiti l'approssimazione di capacità indipendente dalla frequenza è valida.

Altro aspetto interessante dal punto di vista pratico: il bilanciamento potrebbe essere ottenuto anche immaginando di usare un condensatore variabile C_3 e una coppia di resistori fissi. Questa implementazione avrebbe degli svantaggi pratici, poiché non è semplice costruire condensatori variabili che funzionino bene (tanti anni fa si usavano condensatori variabili ad aria negli apparecchi radio fatti con settori semicircolari rotanti che fungevano da armature) e possano essere calibrati in maniera affidabile. Dunque il ponte di de Sauty permette di superare questa difficoltà usando, invece, un potenziometro, che è molto più semplice da costruire e può essere calibrato in modo piuttosto affidabile.

C. Il ponte di Maxwell

Vediamo ora cosa succede se si vuole utilizzare un ponte per misurare un'induttanza. Supponiamo allora che il componente 1 sia un induttore. Se fosse possibile costruire un induttore che ha un'impedenza puramente immaginaria (detta anche reattanza), il problema sarebbe formalmente analogo a quello dei condensatori e l'interesse concettuale scemerebbe. Purtroppo un induttore ha sempre anche una componente reale della sua impedenza, dovuta alla resistenza ohmica del filo che lo costituisce. Raramente si verifica che tale resistenza sia trascurabile, magari a frequenze alte, dove però generalmente il valore della resistenza interna di un induttore aumenta, ad esempio per l'effetto di prossimità, come già abbiamo incontrato. Dato che anche la resistenza interna di un induttore incognito è incognita, è evidente che in questo caso bisogna escogitare un qualche sistema che permetta di eseguire due misure (R_1 e L_1) invece che una sola.

Prima di arrivare a quello che si definisce generalmente (noi definiamo) ponte di Maxwell vero e proprio, vediamo qualche possibilità alternativa, mettendo in luce i limiti pratici o concettuali che vi sono coinvolti. Ricordiamo che per l'induttore (reale) da misurare è $Z_1 = j\omega L_1 + R_1$. I casi di cui si parla qui nel seguito sono rappresentati in Fig. 2, che riporta un'illustrazione adattata dalle trasparenze della Prof. Andreozzi (anno accademico 2010/11).

1. Caso 1

Immaginiamo qui: $Z_2 = j\omega L_2 + R_2$, $Z_3 = R_3$, $Z_4 = R_4$, quest'ultima supposta variabile (il componente 2 è un induttore noto, il 3 e il 4 sono dei resistori, l'ultimo dei quali variabile per permettere il bilanciamento).

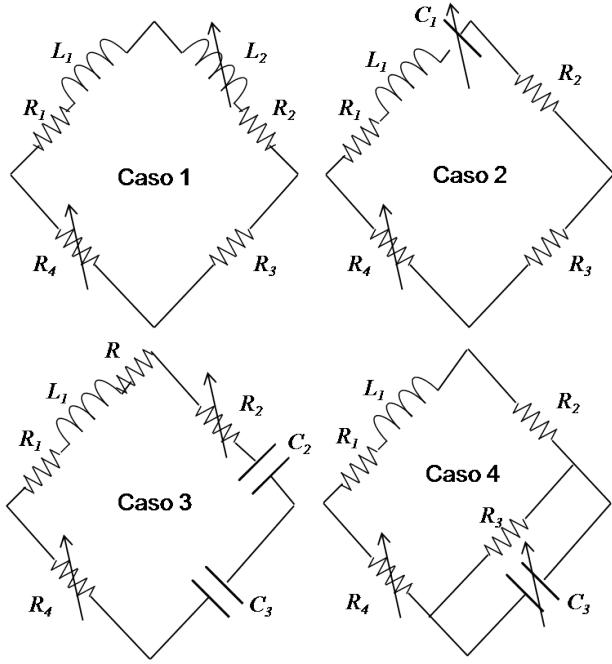


Figura 2. I quattro casi, ovvero configurazioni, discussi nel testo. Il caso 4 è quello che noi indichiamo generalmente come ponte di Maxwell.

Per il bilanciamento si deve verificare: $(j\omega L_1 + R_1)R_3 = (j\omega L_2 + R_2)R_4$, cioè, esaminando separatamente le componenti immaginarie e reali: $L_1 = L_2 R_4 / R_3$ e $R_1 = R_2 R_4 / R_3$. Queste due condizioni possono essere esplorate separatamente, usando rispettivamente un generatore alternato o uno continuo. Ma c'è un problema pratico non da poco: i rapporti L_1/L_2 e R_1/R_2 devono essere uguali. Ora R_1 è supposto incognito tanto quanto L_1 , e quindi non è possibile costruire un induttore di riferimento che abbia L_2 e R_2 tali da soddisfare sempre e comunque la condizione. Possibilità alternativa è quella di supporre anche L_2 variabile (e magari R_2 fissa, oppure variabile indipendentemente da L_2). Si può fare, ma è piuttosto complicato (pensate a come si può realizzare un componente con induttanza variabile) ed è difficile che il tutto funzioni bene. Inoltre, per un fenomeno che già conoscete, gli induttori tendono a “parlare” tra di loro attraverso la mutua induzione, cioè il loro comportamento dipende anche dalla presenza, a breve distanza, degli altri induttori, per cui lo strumento rischia di essere perturbato in maniera dipendente dalle condizioni di uso.

Questi motivi sono sufficienti a decretare la scarsa funzionalità di questa configurazione.

2. Caso 2

Supponiamo qui di “compensare” la reattanza L_1 mettendo in serie al componente 1 un condensatore variabile di capacità C_1 (supponiamo di poter disporre di

un buon condensatore variabile). Allora immaginiamo $Z_1 = j\omega L_1 + R_1 + 1/(j\omega C_1)$. Scegliamo poi tutti gli altri componenti come resistivi, cioè $Z_2 = R_2$, $Z_3 = R_3$, $Z_4 = R_4$, supposta variabile.

Le condizioni di bilanciamento impongono: $(j\omega L_1 + R_1 + 1/(j\omega C_1))R_3 = R_2 R_4$, ovvero, separando le componenti immaginarie e reali: $R_1 = R_2 R_4 / R_3$ e $L_1 = 1/(\omega^2 C_1)$. Quest'ultima espressione ci mostra che la misura di L_1 richiederebbe di conoscere ω , cioè sarebbe funzione di ω , cosa contraria allo spirito della misura a ponte, almeno nella filosofia convenzionale di queste misure. Infatti l'uso pratico dei ponti di misura risale a un'epoca nella quale la misura della frequenza, come anche la generazione di segnali a una data frequenza, era poco agevole, non essendo disponibile la tecnologia digitale che è alla base del funzionamento dei moderni frequenzimetri, e anche dei moderni generatori di funzione. Inoltre ci sarebbe un altro problema: la presenza del condensatore C_1 rende impossibile utilizzare il ponte in continua (non passerebbe corrente lungo il ramo A), per cui le due condizioni di cui sopra non possono essere verificate indipendentemente. Quindi questa configurazione è assai poco funzionale e va scartata.

3. Caso 3

Supponiamo qui di collegare in serie all'induttore reale incognito una resistenza ulteriore, che chiamiamo R . Inoltre il componente 2 è la serie di un condensatore (C_2) e una resistenza (supposta variabile e indicata come R_2), mentre componenti 3 e 4 sono rispettivamente un condensatore C_3 e una resistenza variabile R_4 . Osservate subito un buon vantaggio pratico della configurazione: gli elementi variabili sono solo resistenze. Notate però anche lo svantaggio, comune al caso 2, consistente nell'impossibilità di usare il ponte in continua (non passerebbe corrente nel ramo B).

Le impedenze sono: $Z_1 = j\omega L_1 + R_1 + R$, $Z_2 = 1/(j\omega C_2) + R_2$, $Z_3 = 1/(j\omega C_3)$ e $Z_4 = R_4$. La condizione di bilanciamento è: $(j\omega L_1 + R_1 + R)/(j\omega C_3) = (R_2 + 1/(j\omega C_2))R_4$, ovvero, separando le componenti: $L_1 = R_4 R_2 C_3$ e $R_1 = R_4(C_3/C_2) - R$. Si ritrova l'indipendenza dalla frequenza, che è cosa buona, ma si incontra un ulteriore problema. Il bilanciamento di tutte e due le componenti deve avvenire in alternata e la determinazione di R_1 richiede di aggiustare la resistenza variabile (eventualmente calibrata) R_4 . Però il valore di R_4 influenza anche la misura di L_1 . Se ci pensate un po', potete facilmente concludere che l'uso di questa configurazione è complicato: avete due manopole, dovete operare sempre in alternata, una manopola, quella di R_4 , serve per misurare R_1 ma serve anche, moltiplicata per R_2 , che è pure variabile, per misurare L_1 . In altre parole, componente reattiva e resistiva dell'induttore 1 non possono essere misurate in modo indipendente tra loro, e ciò ci induce a scartare anche questa configurazione.

4. Caso 4 - Ponte di Maxwell

In questa configurazione la principale particolarità che salta all'occhio è che il componente 3 è costituito da un parallelo tra un condensatore (variabile) C_3 e una resistenza R_3 . Come nel caso 1, il ponte che si costruisce può operare sia in continua che in alternata. Inoltre questo caso ha il vantaggio pratico rispetto al caso 1 dovuto alla circostanza che c'è un solo induttore (quindi la mutua induzione non disturba).

Le impedenze sono: $Z_1 = j\omega L_1 + R_1$, $Z_2 = R_2$, $Z_3 = R_3/(1 + j\omega R_3 C_3)$ (resistore e condensatore sono in pa-

rallelo tra loro), $Z_4 = R_4$, supposta variabile. Il bilanciamento richiede: $(j\omega L_1 + R_1)(R_3/(1 + j\omega R_3 C_3)) = R_2 R_4$. Separando le componenti immaginaria e reale, a cui si può accedere usando un generatore alternato o continuo, si ha: $L_1 = R_4 C_3 R_2$ e $R_1 = R_2 R_4 / R_3$.

L'indipendenza dalla frequenza è soddisfatta. Inoltre è vero che anche qui le manopole sono due ed è il loro prodotto a determinare le misure di interesse, ma stavolta si può pensare di aggiustare prima una (per esempio R_4) e poi l'altra (per esempio C_2) manopola bilanciando prima in continua e poi in alternata, o viceversa.

Tutto questo rende(va) molto funzionale la configurazione, al punto di farle meritare il nome di ponte di Maxwell.

Esercizi sulla trasformata di Fourier (FFT)

francesco.fuso@unipi.it

(Dated: version 2 - FF, 5 aprile 2018)

Questa nota ha lo scopo di introdurre l'argomento dell'*analisi di Fourier* condotta numericamente (*Discrete Fourier Transform* - DFT) attraverso il metodo della *Fast Fourier Transform* (FFT) applicato a dati acquisiti con Arduino durante le esercitazioni pratiche. Dal punto di vista concettuale, si tratta di un argomento potentissimo e di importanza ubiqua in tantissimi settori della ricerca scientifica. Il nostro approccio al problema è minimale e ci accontentiamo di impiegare una versione dell'algoritmo FFT implementato in Python per trattare dati di origine sperimentale. La nota, dopo una breve introduzione, fornisce suggerimenti pratici e istruzioni per lo svolgimento dell'esercizio.

I. ANALISI DI FOURIER

In termini molto generali e volutamente privi dei contenuti concettuali e matematici che avete trovato, o troverete, in altri corsi, l'analisi di Fourier può essere considerata un meraviglioso strumento per esaminare sistemi fisici, in particolare quelli che si comportano in modo "lineare", in cui, per esempio, i segnali in "uscita" e "ingresso" sono legati tra loro da una relazione causale che può essere modellata da una funzione (semplice e possibilmente analitica) di trasferimento.

Questa funzione, generalmente complessa, descrive il comportamento del sistema in un dato dominio, che è molto spesso quello delle frequenze, o frequenze angolari, come abbiamo più volte avuto modo di incontrare nel nostro corso. La *trasformata* di Fourier, che è alla base del metodo di analisi che vogliamo trattare, consente di passare al dominio "coniugato", che è quello dei tempi. Naturalmente esistono altri domini coniugati tra loro, cioè tali che il prodotto tra le variabili (coordinate) dei due domini sia adimensionale; tuttavia in questa nota, in accordo con le nostre esigenze e possibilità, ci restringiamo a esaminare la trasformazione dal dominio dei tempi a quello delle frequenze. L'operazione inversa, che può anche essere eseguita, è chiamata *anti-trasformata*: nei nostri esempi essa permette di passare dal dominio delle frequenze a quello dei tempi.

Per essere estremamente grossolani e animati da senso pratico, possiamo vedere la trasformata di Fourier come un'estensione del concetto di *serie* di Fourier che già abbiamo incontrato e utilizzato nel nostro corso. Abbiamo in particolare già mostrato come una generica funzione *periodica* di frequenza angolare ω possa essere espressa come somma di (tante) funzioni seno e coseno, dette armoniche, ognuna di frequenza angolare $k\omega$, con k intero. I pesi della somma li abbiamo definiti *coefficienti* dello sviluppo in serie di Fourier, e, all'epoca, abbiamo anche stabilito una regola per la loro determinazione.

La trasformata di Fourier rappresenta, in pratica, l'applicazione dello stesso concetto a funzioni *non periodiche* ma dipendenti dal tempo secondo un andamento "qualsiasi". Detta $g(t)$ una tale funzione, si ha che la sua trasformata di Fourier è, a parte eventuali coefficienti di

normalizzazione,

$$\mathcal{F}\{g(t)\} = \tilde{g}(f) = \int_{-\infty}^{\infty} g(t) \exp(j2\pi ft) dt , \quad (1)$$

dove $j = \sqrt{-1}$ è l'unità immaginaria e con $f = \omega/(2\pi)$ indichiamo una frequenza. La $\tilde{g}(f)$, calcolata per una data f (ovvero per l'intervallo infinitesimo $f, f+df$, secondo quanto discuteremo nel seguito), rappresenta in sostanza il valore dei coefficienti (complessi) dell'espansione di Fourier della $g(t)$.

L'anti-trasformata può essere definita come

$$\mathcal{F}^{-1}\{\tilde{g}(f)\} = g(t) = \int_{-\infty}^{\infty} \tilde{g}(f) \exp(-j2\pi ft) df , \quad (2)$$

sempre a parte eventuali coefficienti di normalizzazione.

La bellezza dell'analisi di Fourier può essere facilmente compresa ricordando quanto abbiamo fatto, per esercizio, con la serie di Fourier. All'epoca abbiamo osservato come il comportamento di un sistema (un filtro passabasso o passa-alto, o una combinazione dei due) per un segnale periodico non sinusoidale $g_{in}(t)$ (un'onda quadra o triangolare) potesse essere ben riprodotto utilizzando la funzione di trasferimento $T(f)$ determinata per segnali sinusoidali, e ricostruendo il segnale di uscita $g_{out}(t)$ come sovrapposizione di tante componenti che rappresentavano l'applicazione della funzione di trasferimento alle componenti del segnale in ingresso. Questa procedura può essere generalizzata, affermando che

$$g_{out}(t) = \mathcal{F}^{-1}\{\tilde{g}_{out}(f)\} \quad (3)$$

$$\tilde{g}_{out}(f) = T(f)\tilde{g}_{in}(f) \quad (4)$$

$$\tilde{g}_{in}(f) = \mathcal{F}\{g_{in}(t)\} , \quad (5)$$

cioè, in sostanza, che il comportamento del sistema nel dominio delle frequenze è descritto da una semplice operazione di prodotto tra funzione di trasferimento e trasformata di Fourier del segnale in ingresso, e che il segnale di uscita può essere ricostruito anti-trasformando tale prodotto (tutto questo, a parte eventuali coefficienti di normalizzazione ha, come sapete, il nome di "prodotto di convoluzione").

In questa nota, e almeno per questo anno accademico, non applicheremo l'analisi di Fourier a questa interessantissima casistica, limitandoci solo a eseguire trasformate

di Fourier $\tilde{g}(f)$ di segnali $g(t)$ dipendenti dal tempo e acquisiti sperimentalmente con Arduino in qualche esperienza. Dunque in questa nota ci proponiamo di costruire lo spettro dei coefficienti di Fourier di alcuni segnali acquisiti in laboratorio: questo è anche il contenuto dell'esercizio obbligatorio proposto.

Anche questa semplice procedura può risultare estremamente utile. Pensate ad esempio ad un segnale periodico, a una certa frequenza, che è sovrapposto a un rumore, periodico a una o più frequenze, o addirittura aperiodico. Lo spettro del segnale consente di individuare le sue componenti a diverse frequenze permettendo, ad esempio, di determinare l'ampiezza dell'oscillazione del segnale periodico anche se questa è praticamente dello stesso ordine, o addirittura minore, rispetto all'ampiezza del rumore. Inoltre il passaggio dal dominio dei tempi a quello delle frequenze può permettere di ricavare la curva di risposta (spettro) di un oscillatore partendo da misure in funzione del tempo, che talvolta sono le uniche a poter essere eseguite. La possibilità di anti-trasformare consente anche l'operazione inversa, cioè costruire l'andamento temporale di un'oscillazione smorzata partendo dai dati spettrali.

Prima di procedere con le istruzioni pratiche, osserviamo che la trasformata di Fourier, Eq. 1, è per sua definizione, complessa. Nelle nostre esperienze pratiche il segnale è generalmente una d.d.p. acquisita da Arduino in funzione del tempo, dunque si tratta di una grandezza reale. Il *modulo* della trasformata di Fourier fornisce già praticamente tutte le informazioni rilevanti sullo spettro delle ampiezze: ad esempio, eventuali picchi nello spettro della $\tilde{g}(f)$ a determinati valori delle f indicano delle periodicità nella $g(t)$ a quelle frequenze.

Talvolta, per esempio nello studio degli oscillatori smorzati, può essere rilevante calcolare il *modulo quadro* $|\tilde{g}(f)|^2$, a cui si dà spesso il nome di *spettro di potenza*. Questa denominazione tiene conto del fatto che la potenza è proporzionale al modulo quadro dell'ampiezza: dunque tale spettro fornisce informazioni sulle componenti spettrali della potenza che interessa il sistema sotto analisi, cioè, nel nostro caso, gli eventuali circuiti realizzati.

Per quanto riguarda le dimensioni e le unità di misura delle grandezze ottenute tramite FFT, esse possono essere ricavate dalla definizione di Eq. 1. Dal punto di vista dimensionale, notate che moltiplicare la $g(t)$ per un tempo, come in Eq. 1, equivale a dividere per una frequenza. Di conseguenza, e in accordo con quanto sopra affermato sul fatto che $\tilde{g}(f)$ rappresenta i coefficienti di Fourier nell'intervallo $f, f + df$, la $\tilde{g}(f)$ può essere considerata come una *densità spettrale di ampiezza*; analogamente la $|\tilde{g}(f)|^2$ una *densità spettrale di potenza*. Nel caso continuo della matematica, l'intervallo di frequenza può essere reso piccolo a piacere, fino a diventare un infinitesimo df . Nella realtà fisica della trasformata discreta, l'intervallo di frequenza Δf è finito e la sua larghezza è determinata dalle caratteristiche del campionamento della $g(t)$, secondo quanto illustreremo nel seguito. Per fare un esem-

pio, supponiamo di avere una funzione perfettamente sinusoidale (monocromatica) di frequenza f_0 determinata. Secondo la definizione di Eq. 1, il suo spettro sarebbe rappresentato da una funzione delta centrata in $f = f_0$. Tuttavia, se la funzione è discreta, cioè campionata su intervalli temporali finiti, la FFT produce un picco di altezza finita e larghezza non nulla, corrispondente proprio alla risoluzione in frequenza Δf .

Il valore numerico di tale picco dipende dalla sua larghezza, cioè da Δf , oltre che dall'ampiezza della funzione sinusoidale ed esistono diverse possibilità di normalizzazione che in questa nota non intendiamo descrivere. Anche in considerazione del fatto che generalmente i nostri segnali campionati sono misurati in unità arbitrarie (i digit di Arduino), per non appesantire troppo l'esercizio va benissimo se vi limitate a esprimere la $\tilde{g}(f)$ ottenuta per FFT in *unità arbitrarie*.

A. DFT e FFT

Dal punto di vista matematico, l'espressione di Eq. 1 è ricchissima di conseguenze e implicazioni, sia nella fisica classica che in quella quantistica (e relativistica). Tuttavia, come al solito, la fisica non è la patria degli integrali, e neanche delle derivate, e il calcolo esplicito dell'integrale che compare nella definizione può essere affrontato solo per casi "modello", quelli che interessano la soluzione di semplici, o semplicissimi, esercizi da libro di testo.

Esistono dei validi strumenti numerici che permettono di eseguire il calcolo in maniera *discreta*, cioè partendo da una $g(t)$ definita per punti, che dunque non è una funzione, ma rappresenta, per esempio, una misura campionata nel tempo (sarebbe più corretto indicarla come $g(t_i)$, ma evitiamo di farlo per non appesantire la presentazione); il risultato della trasformazione è anche uno spettro discreto (che dovrebbe essere indicato con $\tilde{g}(f_j)$), generalmente complesso. Questi strumenti, uno dei quali è appunto la FFT, sono presenti in tutti i software di analisi dati, compreso, ovviamente (e con diverse implementazioni che qui non discuteremo), Python. Se siete interessati, potete sicuramente trovare in rete, o altrove, informazioni su come funzionano gli algoritmi per la DFT e verificare come l'aggettivo fast della FFT sia giustificato da un metodo di calcolo molto efficiente in termini di risorse.

Prima di procedere, diamo qualche chiarimento sull'implementazione FFT in Python a cui facciamo esplicito riferimento. Il problema principale da risolvere riguarda la *scala* di f , cioè quanto valgono il minimo e il massimo valore delle f da usare come asse orizzontale per lo spettro. Infatti, se l'Eq. 1 non dà alcuna indicazione sulla presenza di limiti in f , il calcolo numerico (discreto) impone delle regole precise.

Senza entrare nei dettagli, le regole principali sono le seguenti:

1. la FFT funziona al meglio, cioè senza la necessità di applicare artifici particolari, se l'array di partenza,

- quello che rappresenta la $g(t)$ (in forma discreta, per punti), è composto da 2^n punti, con n intero. Vedete che, spesso, questo requisito è automaticamente soddisfatto quando i dati di partenza sono quelli acquisiti con una delle varie combinazioni di sketch e script che abbiamo impiegato con Arduino nel corso dell'anno.
2. Nell'implementazione da noi usata, l'array trasformato è fatto da $2^{n-1} + 1$ punti. Supponendo array di partenza costituiti da $2^{11} = 2048$ punti: la nostra FFT è rappresentata da array reali di 1025 punti (la metà più uno dei punti di partenza). Il punto “in più”, quello che rende dispari questo valore, si riferisce a $|\tilde{g}(f=0)|$ e, per i nostri scopi, ha scarsa importanza. Esso infatti rappresenta il valore medio del segnale, calcolato nella durata complessiva dell'acquisizione e influenzato, ad esempio, da eventuali offset, bias, etc.. Osservate che, per come è definita la trasformata di Fourier, anche un piccolo offset, o bias, presente nel segnale può condurre a un valore molto alto del coefficiente spettrale per $f = 0$. Quindi non stupitevi se i vostri spettri presenteranno un grosso picco per la componente continua. Ancora, per migliorare la leggibilità dello spettro spesso può essere opportuno rappresentarli con l'asse verticale logaritmico, scelta che in genere permette di visualizzare nello stesso grafico dati che coprono un vasto intervallo di valori.
 3. Lo spettro di Fourier ottenuto parte sempre da $f = 0$ e si estende fino a $f_{max} = 1/(2\Delta t_{eff})$, dove, per noi, Δt_{eff} rappresenta l'intervallo temporale effettivo tra due campionamenti successivi.
 4. L'affermazione precedente accende un campanello di allarme: usando Arduino, l'intervallo Δt_{eff} è normalmente definito con un'accuratezza piuttosto limitata a causa dei (noti) limiti nella definizione e nella misura dei tempi intrinseci nel funzionamento del microcontroller. Nella sua implementazione ordinaria, l'algoritmo FFT richiederebbe dati equispaziati temporalmente, per cui l'uso di Arduino è inevitabilmente accompagnato da una incertezza, difficilmente valutabile, nella costruzione dello spettro.
 5. La risoluzione in frequenza dello spettro ottenuto, cioè la distanza Δf (in unità di f) tra due punti consecutivi dello spettro, è ovviamente pari a $f_{max}/(2^{n-1})$. Poiché l'intervallo complessivo di durata dell'acquisizione è $\Delta T \simeq 2^n \Delta t_{eff}$, dove il simbolo \simeq può essere sostituito per i nostri scopi dal simbolo $=$ (vero se, in pratica, trascuriamo le incertezze nella misura dei tempi di cui sopra), si ha anche $\Delta f = 1/\Delta T$. Nella pratica conviene servirsi proprio della misura di ΔT per stabilire la scala in frequenza dei nostri spettri, assumendo implicitamente che il campionamento sia equispaziato temporalmente.
 6. Sulla base di quanto appena affermato, è evidente che la risoluzione in frequenza migliora all'aumentare della durata temporale del record di dati. Visto che nelle nostre acquisizioni operiamo tipicamente con un intervallo nominale di campionamento relativamente breve (decine o centinaia di μs , in genere), costruire un record di durata sufficiente per garantire una ragionevole risoluzione in frequenza implica acquisizioni di record relativamente lunghi, cioè costituiti da un numero relativamente grande di dati. Come sappiamo, questo può essere facilmente realizzato (con segnali ripetitivi) usando opportune strategie di sincronizzazione per Arduino. Naturalmente anche in questo caso non ha senso esagerare: pretendere una risoluzione in frequenza molto elevata cozza con l'incertezza nella determinazione dei tempi tipica di Arduino. Inoltre, in particolare nei casi in cui ci si aspetta la presenza di picchi relativamente “larghi” nello spettro, per esempio per gli oscillatori *RLC* fortemente smorzati, può essere sufficiente anche esaminare acquisizioni con un ridotto numero di punti, fino a soli 256 punti.
 7. In generale, non esiste un criterio che permetta di aggiustare a priori il rate di campionamento, ovvero l'intervallo Δt_{nom} , allo scopo di ottenere la “migliore” FFT. L'applicazione di un famoso teorema, detto *di Nyquist*, suggerisce come nel caso di funzioni periodiche, che sono quelle di nostro interesse in questo semplice esercizio, bastino tre punti per ciclo per ottenere informazioni rilevanti sulle periodicità. Naturalmente il consiglio è quello di campionare almeno una decina di punti per ogni ciclo. Notate che forti sovra-campionamenti, cioè avere un gran numero di punti per ogni ciclo, unito alla lunghezza relativamente piccola dei record acquisiti, può condurre ad avere spettri poco risolti in frequenza. Viceversa, forti sotto-campionamenti, in cui solo pochi punti sono acquisiti per ogni ciclo, può condurre ad artefatti di diverso tipo, il più frequente dei quali consiste nell'ottenere picchi spettrali “frastagliati”.
 8. Per un utente esperto sono a disposizione diverse strategie finalizzate a controllare numericamente, entro certi limiti, l'esito della FFT. La più importante riguarda l'impostazione della cosiddetta “finestra”. La funzione finestra rappresenta in pratica il “peso” che viene attribuito ai vari punti che costituiscono la $g(t)$ nell'operazione numerica di trasformata. Nella configurazione più semplice, quella consigliata per lo svolgimento di questo esercizio, la finestra è piatta (“rettangolare”), cioè tutti i punti vengono considerati con lo stesso peso. È probabile che in futuro vi troverete a usare finestre di altro genere, che possono essere utili per ridurre la comparsa di eventuali artefatti nello spettro.

Dal punto di vista pratico, poiché le grandezze campionate sono, nel nostro caso, *reali* e il nostro interesse è quello di determinare il *modulo* della trasformata di Fourier, il comando di Python che possiamo usare è `abs(numpy.fft.rfft(V))`, dove V rappresenta l'array di dati della nostra misura (il segnale) e `abs` indica l'operazione di calcolo del modulo dell'array complesso prodotto dalla FFT. Questo comando genera un nuovo array, che possiamo graficare in funzione di un array, da costruire a parte, che riporta i valori discreti della frequenza f determinati come sopra specificato. Il grafico ottenuto rappresenta lo spettro \tilde{V} del segnale.

Se, invece, foste interessati a costruire lo spettro in potenza, allora il comando da impiegare sarebbe `(abs(numpy.fft.rfft(V)))**2`, che, di nuovo, produrrebbe un array da graficare in funzione dell'array f costruito come appena illustrato.

II. L'ESERCIZIO

All'esercizio si può rispondere con qualche grafico e qualche riga di commento (da caricare in formato pdf, possibilmente in unico file, sulla pagina di e-learning del corso). L'esercizio richiede di calcolare tramite FFT gli spettri di qualche acquisizione realizzata con Arduino nel corso delle esperienze pratiche. Per i motivi prima esposti, si consiglia di impiegare acquisizioni con record lunghi, per esempio quelle create dalla combinazione di sketch e script `synclong2016`, `harmlong` e `transosc`. Normalmente queste combinazioni creano record composti da 2048 coppie di dati, tempo (in μs) e valore digitalizzato (in unità arbitrarie, ovvero, nel nostro linguaggio, digit): la trasformata di Fourier di questi dati dà luogo ad un array composto da 1025 punti, compreso lo zero. Tuttavia in alcuni casi, per esempio per lo studio degli oscillatori *RLC*, può essere sufficiente esaminare record da soli 256 punti.

L'esercizio può essere compiuto su qualsiasi file in vostro possesso che rappresenti l'andamento temporale di un segnale, con caratteristiche preferibilmente note e riferibili a precise situazioni fisiche di interesse nel corso (da specificare nei commenti).

Qualche esempio di dati che potrebbero essere impiegati è elencato qui nel seguito.

A. Forme d'onda

Nelle prime esperienze in cui abbiamo usato Arduino ci siamo posti il problema di verificare la ricostruzione dei segnali periodici prodotti dal generatore di forme d'onda. Noi già conosciamo il metodo, quello della serie di Fourier, adatto per determinare i coefficienti dell'espansione per forme d'onda "semplici". Come ricordate, in un precedente esercizio abbiamo proprio usato la serie di Fourier per ricostruire numericamente delle forme d'onda periodiche. Qui la finalità dell'esercizio è diversa: si

parte da dati "reali", necessariamente fisici, cioè dotati di "imperfezioni" dovute al generatore o al processo di campionamento/digitalizzazione, e di questi dati si calcola la FFT.

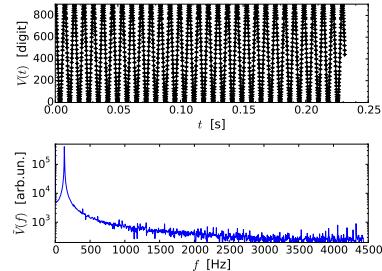


Figura 1. Segnale $V(t)$ e spettro FFT, $\tilde{V}(f)$, corrispondente per una forma d'onda sinusoidale prodotta dal generatore di forme d'onda e acquisita in record lunghi da 2048 punti. Il segnale è rappresentato con punti, dotati delle barre di errore convenzionali, raccordati da una linea continua, che serve solo da guida per gli occhi. Lo spettro è rappresentato in unità arbitrarie e in scala logaritmica: per lo spettro non è riportata alcuna barra di errore, che sarebbe complicato (e poco opportuno) determinare, e il grafico è realizzato impiegando una linea continua.

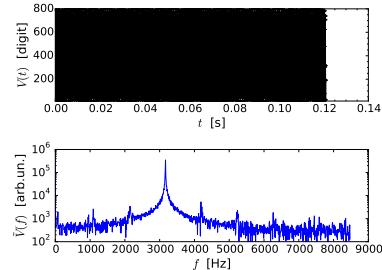


Figura 2. Analogico di Fig. 1 per un'altra forma d'onda sinusoidale, evidentemente acquisita in modo da registrare un grande numero di cicli (l'andamento della $V(t)$ nel pannello superiore risulta completamente non intelligibile). Notate l'aumento nella risoluzione in frequenza dello spettro.

Per intenderci, se avete un record che rappresenta la digitalizzazione di un'onda quadra o triangolare, la FFT dovrebbe mostrare un picco principale alla frequenza dell'onda stessa (a parte il contributo della frequenza zero, che corrisponde alle componenti continue di offset o bias, e che, per i nostri scopi, non è rilevante), più altri picchi minori che corrispondono alle diverse armoniche. La frequenza di questi picchi dovrebbe essere nel corretto rapporto con la frequenza fondamentale e anche la loro ampiezza dovrebbe soddisfare le aspettative (ripensate all'espansione in serie di seni e coseni per queste forme d'onda: vedrete che le armoniche presenti dovrebbero essere solo le dispari e che i coefficienti dovrebbero seguire un andamento specifico per la forma d'onda). Aveste anche un'acquisizione di "pinna di squalo" con un numero

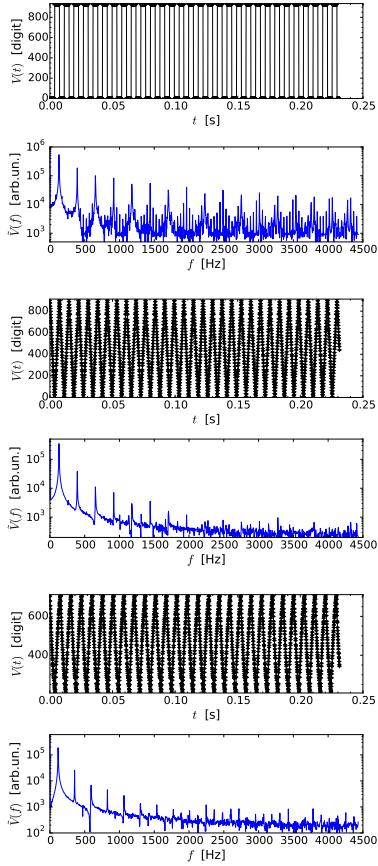


Figura 3. Analogo di Fig. 1 per una forma d’onda quadra, una triangolare e una “pinna di squalo” acquisita nell’esperienza pratica con il circuito RC .

sufficiente di punti, potrebbe anche essere interessante studiarne la trasformata di Fourier.

Invece l’acquisizione di una forma d’onda sinusoidale dovrebbe mostrare solo un picco spettrale, alla frequenza dell’onda stessa. In linea di principio, questo picco dovrebbe apparire “strettissimo” (in matematica una “delta”), a testimonianza del carattere monocromatico della forma d’onda sinusoidale.

I condizionali che ho volutamente impiegato stanno a ricordare che la realtà sperimentale è, appunto, una realtà: infatti potrebbero essere presenti dei rumori sovrapposti al segnale (tipica è la presenza di rumori attorno alla frequenza di rete o suoi multipli), oppure potrebbero diventare rilevanti eventuali effetti legati al campionamento (fluttuazioni nell’intervallo di campionamento, errori sistematici nella digitalizzazione da parte di Arduino, etc.). Infine, poiché le forme d’onda sono anch’esse “reali”, cioè prodotte da un dispositivo reale quale il generatore di forme d’onda, esse potrebbero presentare differenze rispetto alle aspettative, per esempio la presenza di armoniche inattese e, soprattutto, una larghezza finita dei picchi spettrali (la monocromaticità prevista dalla matematica non è mai possibile nella realtà).

Le Figs. 1-3 presentano alcuni esempi, che sono riportati qui senza commento (nel vostro esercizio mi aspetto qualche commento in più, almeno sui parametri di acquisizione, che devono essere specificati, e sulla frequenza della forma d’onda acquisita). Il segnale campionato è denominato $V(t)$ e rappresentato in unità arbitrarie di digitalizzazione (digit).

B. Oscillatore smorzato

Un valido esercizio per apprezzare la bellezza dell’analisi di Fourier è costituito dalla trasformazione del segnale acquisito con l’oscillatore armonico smorzato RLC . Nel nostro corso analizziamo il comportamento di questo circuito prima nel dominio dei tempi, ottenendo dei segnali rappresentativi di un’oscillazione smorzata, e poi nel dominio delle frequenze, ricostruendo la funzione di trasferimento del circuito e analizzando il suo andamento al variare della frequenza.

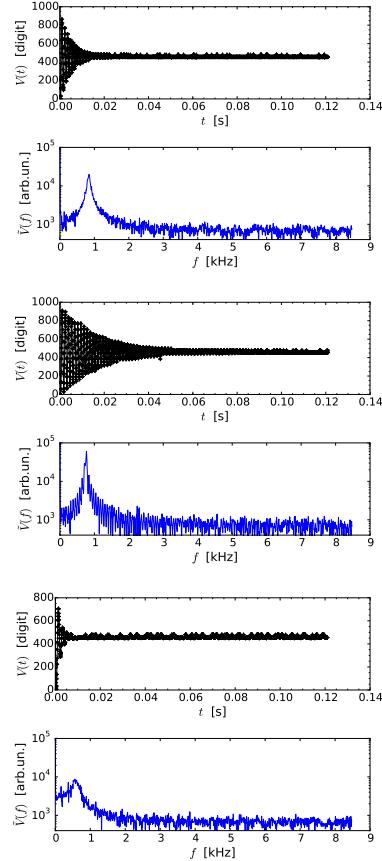


Figura 4. Analogo di Fig. 1 per segnali acquisiti nelle esperienze con l’oscillatore smorzato RLC operato in varie condizioni.

La trasformata di Fourier permette di passare da un dominio all’altro. Dunque, almeno tralasciando le “imperfezioni” dovute al campionamento, il modulo dello

spettro FFT del segnale acquisito in determinate condizioni deve riprodurre l'andamento del guadagno, o attenuazione, nelle stesse condizioni. Potete facilmente verificarlo, osservando in particolare come lo spettro ottenuto sia piccato attorno alla frequenza di risonanza e come la sua larghezza dipenda dal fattore di qualità, ovvero dallo smorzamento dell'oscillatore: la larghezza del "picco di risonanza" nello spettro è tanto maggiore quanto più smorzato è l'oscillatore.

Anche in questo caso si riportano alcuni esempi, volutamente senza commento (ma nei vostri esercizi mi aspetto commenti e qualche indicazione sui parametri sperimentali), nelle Figs. 4, 5. Notate che in alcuni casi sono state impiegate delle acquisizioni costituite da soli 256 punti, con ovvie conseguenze sulla risoluzione in frequenza dello spettro.

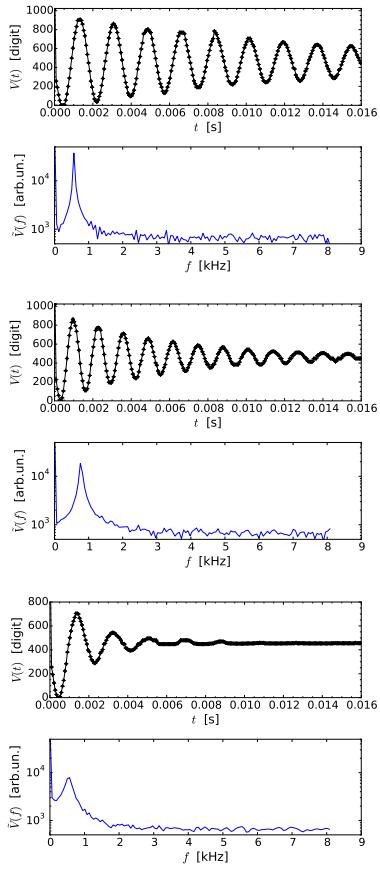


Figura 5. Analogo di Fig. 4: gli spettri sono qui costruiti a partire da acquisizioni composte da soli 256 punti, con ovvia perdita di risoluzione in frequenza negli spettri.

C. Oscillatore a reazione (con transistor)

Infine, un'altra interessante occasione di impiego della FFT è con l'oscillatore a reazione basato su transistor BJT a emettitore comune e rete di sfasamento per ottenere un feedback positivo. In queste condizioni la d.d.p. oscillante che si ottiene in uscita è attesa avere una forma "strana", che assomiglia solo lontanamente a una sinusoida a causa dei meccanismi di operazione dell'oscillatore. Di tutto questo è fatto cenno in un'altra nota.

In questa sede ci limitiamo a presentare in Fig. 6 alcuni esempi acquisiti in diverse condizioni di operazione. Anche qui rinunciamo a specificare commenti (che però dovrebbero essere presenti, assieme alla descrizione dei parametri sperimentali, nel vostro esercizio).

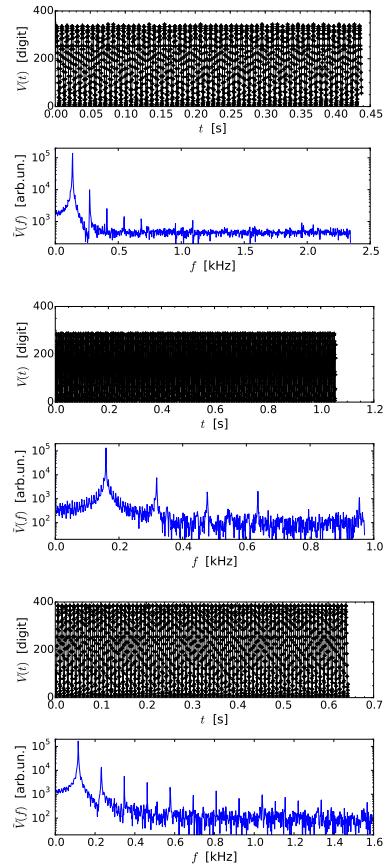


Figura 6. Analogo di Fig. 1 per segnali acquisiti in oscillatori a reazione (con transistor) operati in diverse condizioni.