Capacità di servizio e tempi di attesa

Indice

Motivazione gestionale

Costruzione guidata del modello

Esempio svolto a mano

Caso di studio e risultati

Analisi di sensitività: il prezzo di una promessa

Esercizi

Obiettivi (1/2)

Alla fine di questo capitolo sarai in grado di:

- usare le formule della coda M/M/1 dentro un problema di ottimizzazione;
- ricavare l'ottimo analiticamente e confermarlo numericamente;

Obiettivi (2/2)

Alla fine di questo capitolo sarai in grado di:

- dare un prezzo a una promessa di servizio;
- proteggere il dimensionamento dall'incertezza sulla domanda.

Prerequisiti

- Dal modulo *Ottimizzazione convessa*, Capitolo 3 «Funzioni convesse e concave»: le funzioni convesse di una variabile e la loro minimizzazione
- Da questo modulo, Capitolo 2 «Richiami: convessità, programmazione quadratica e condizioni KKT»: i programmi quadratici convessi, le condizioni KKT e i moltiplicatori come prezzi ombra
- Da questo modulo, Capitolo 3 «Il solver per i modelli non lineari»: i vincoli funzionali e la stima dei prezzi ombra per perturbazione

Motivazione gestionale

Introduzione: Motivazione gestionale

- In questo capitolo decidiamo **quanta capacità** assegnare a un servizio
- Classe del problema: **NLP convesso** (coda M/M/1)
- La teoria delle code dà la relazione quantitativa tra capacità e attesa
- L'ottimizzazione la trasforma in una **decisione**

Capacità contro attesa

- Quanta capacità assegnare a un servizio clienti, a uno sportello, a un servizio cloud?
- Più capacità costa; poca capacità fa **esplodere le attese**
- Ogni servizio con arrivi aleatori (call center, pronto soccorso, help desk, API di un servizio cloud) vive lo stesso dilemma: la capacità costa, l'attesa pure
- La teoria delle code dà la relazione esatta tra le due; l'ottimizzazione sceglie il punto giusto

Il messaggio centrale

Il messaggio è contro-intuitivo per chi ragiona "a efficienza":

l'utilizzazione ottima non è il 100%

Vale da solo una discussione in consiglio di amministrazione.

Riepilogo: Motivazione gestionale

- Capacità e attesa sono due costi in conflitto
- La teoria delle code li lega con formule esatte
- L'utilizzazione ottima **non** è il 100%

Costruzione guidata del modello

Introduzione: Costruzione guidata del modello

- Descriviamo il problema a parole e fissiamo i dati
- Introduciamo la variabile decisionale: la capacità μ
- Scriviamo il modello di costo totale e lo risolviamo in **forma chiusa**

Il problema a parole

- *Decidiamo* quanta capacità di servizio acquistare
- *L'obiettivo* è minimizzare il costo totale: capacità comprata più il valore del tempo che i clienti passano nel sistema
- *Il vincolo* strutturale è uno solo: la stabilità della coda ($\mu > \lambda$)
- Nelle varianti si aggiunge la promessa di servizio $w \leq w_{\max}$

La coda M/M/1

- Usiamo il modello di coda più semplice, in notazione standard **M/M/1**:
  - arrivi *markoviani* (processo di Poisson)
  - tempi di servizio *markoviani* (esponenziali)
  - *un* servente
- Per questo sistema valgono formule chiuse
- Ciò che qui è nuovo: trattare la capacità μ come una **variabile di ottimizzazione**

Dati (input del modello)

Simbolo	Tipo	Significato
λ	$\in \mathbb{Q}_{>0}$	tasso medio di arrivo dei clienti (richieste/ora)
c	$\in \mathbb{Q}_{>0}$	costo di un'unità di capacità (€ per unità di μ all'ora)
h	$\in \mathbb{Q}_{>0}$	valore di un'ora di permanenza di un cliente nel sistema (€)

Variabile decisionale e grandezze derivate

- Introduciamo una variabile continua:

  $\mu =$ capacità del servizio (tasso di servizio, richieste/ora)

  con il requisito di stabilità $\mu > \lambda$

- Tempo medio nel sistema (in letteratura: W):

$$w(\mu) = \frac{1}{\mu - \lambda}$$

- Numero medio di clienti nel sistema (legge di Little; in letteratura: L):

$$l(\mu) = \frac{\lambda}{\mu - \lambda}$$

Il modello

Usando questa variabile, il modello per il problema è il seguente:

Modello di costo totale capacità + attesa

$$\min \quad c\,\mu + h\,\frac{\lambda}{\mu - \lambda} \tag{11.1a}$$

$$\text{soggetto a} \quad \mu > \lambda. \tag{11.1b}$$

Funzione obiettivo e vincolo

- La funzione obiettivo è la somma di due termini:
  - $c\,\mu$ compra la capacità
  - $h \cdot l(\mu)$ monetizza il tempo complessivo che i clienti passano nel sistema
- L'obiettivo è **convesso** su $(\lambda, +\infty)$
- Il vincolo $\mu > \lambda$ definisce la variabile e garantisce la stabilità della coda: con $\mu \leq \lambda$ la fila cresce senza limite

Soluzione in forma chiusa

Annullando la derivata dell'obiettivo si ottiene:

$$C'(\mu) = c - \frac{h\,\lambda}{(\mu - \lambda)^2} = 0$$

$$\Longrightarrow \qquad \boxed{\mu^* = \lambda + \sqrt{h\,\lambda / c}}$$

Capacità = domanda + scorta di sicurezza

- La formula ha la struttura classica "capacità = domanda + scorta di sicurezza"
- Si copre la domanda media λ e si aggiunge un margine $\sqrt{h\lambda/c}$
- Il margine cresce con il valore del tempo dei clienti e decresce con il costo della capacità
- Il margine cresce come $\sqrt{\lambda}$: i sistemi grandi godono di economie di scala nel "cuscinetto" (**pooling**)

Riepilogo: Costruzione guidata del modello

- Una sola variabile continua: la capacità $\mu > \lambda$
- Costo totale $c\mu + h\lambda/(\mu - \lambda)$, convesso
- Ottimo in forma chiusa: $\mu^* = \lambda + \sqrt{h\lambda/c}$
- Il margine di sicurezza cresce come $\sqrt{\lambda}$ (pooling)

Esempio svolto a mano

Introduzione: Esempio svolto a mano

- Applichiamo la formula di μ^* a un piccolo sportello
- Calcoliamo utilizzazione, tempo medio e costo all'ottimo
- Vediamo perché "tagliare la capacità" è una cattiva idea

Esempio: sportello con 8 clienti l'ora (1/2)

Dati: $\lambda = 8$/h, $c = 2$ € per unità di capacità, $h = 4$ €/h.

- **Passo 1.**

$$\mu^* = 8 + \sqrt{4 \cdot 8/2} = 8 + 4 = 12 \text{ clienti/ora}$$

- **Passo 2.** Utilizzazione $\rho = 8/12 = 66{,}7\%$
- Tempo medio $w = 1/(12 - 8) = 15$ minuti
- Costo $C = 2 \cdot 12 + 4 \cdot 8/4 = 32$ €/h

Esempio: sportello con 8 clienti l'ora (2/2)

- **Passo 3 — perché non tagliare la capacità?**
- Con $\mu = 9$ ($\rho = 89\%$, "più efficiente"):

$$C = 18 + 32/1 = 50 \text{ €/h}$$

  molto peggio
- L'apparente spreco (33% di capacità inutilizzata) è ciò che tiene corte le code

Riepilogo: Esempio svolto a mano

- All'ottimo: $\mu^* = 12$, $\rho = 66{,}7\%$, $w = 15$ minuti, $C = 32$ €/h
- Aumentare l'utilizzazione all'89% fa salire il costo a 50 €/h
- La capacità inutilizzata non è spreco: tiene corte le code

Caso di studio e risultati

Introduzione: Caso di studio e risultati

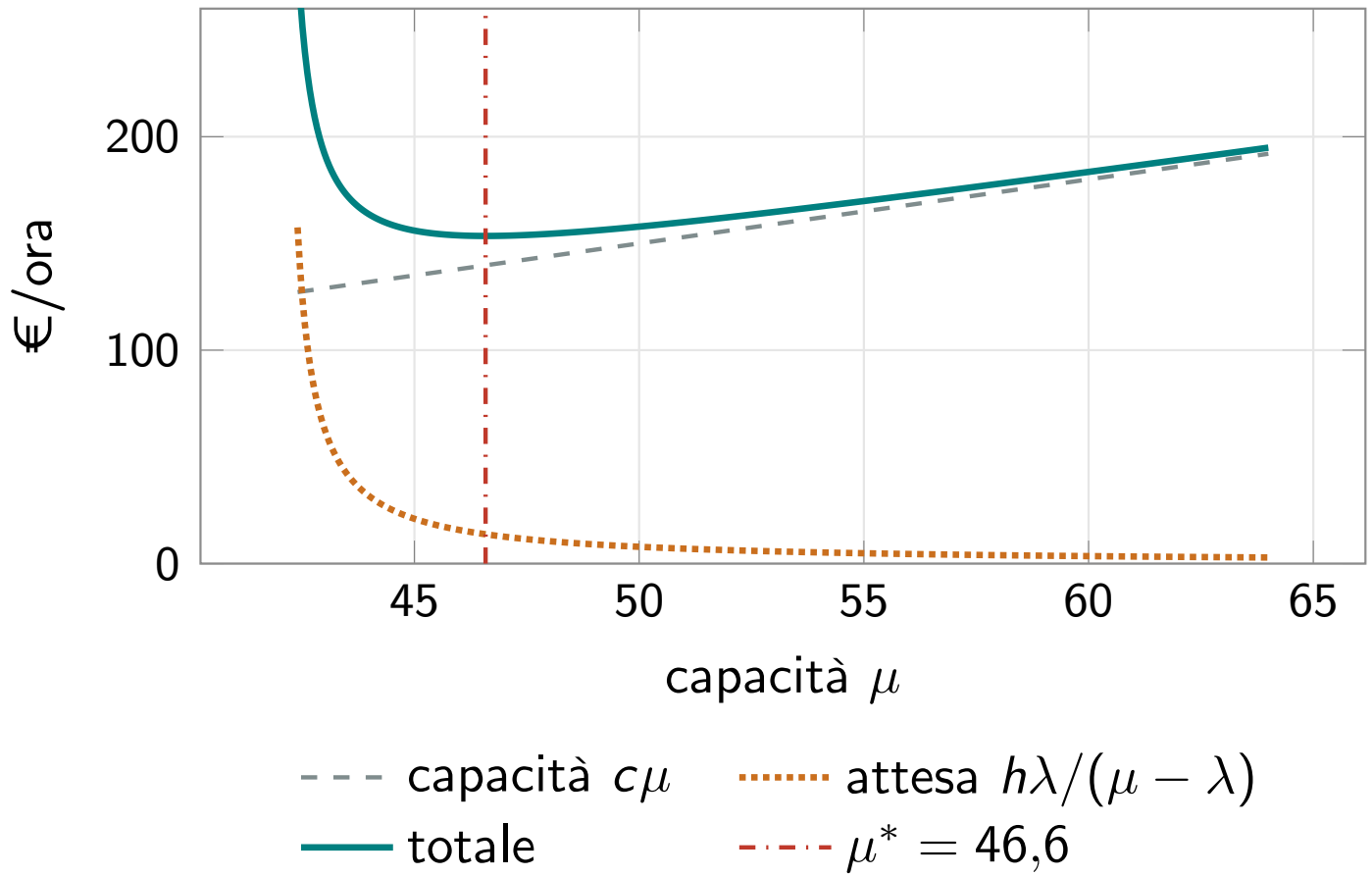
- Un servizio clienti: $\lambda = 42$ richieste/ora, $c = 3$ €, $h = 1{,}5$ €
- Confrontiamo la soluzione analitica con quella di Gurobi
- Guardiamo i due grafici chiave: costo totale e "muro dell'utilizzazione"

Risultati

Servizio clienti: $\lambda = 42$ richieste/ora, $c = 3$ €, $h = 1{,}5$ €.

```
Analitico : mu* = 42 + sqrt(1,5*42/3) = 46,583
            costo 153,495 EUR/h
Gurobi    : mu* = 46,582 (bilineare w(mu-lam) = 1,
            NonConvex=2: coincide)
All'ottimo: rho = 90,2%   w = 13,1 minuti
```

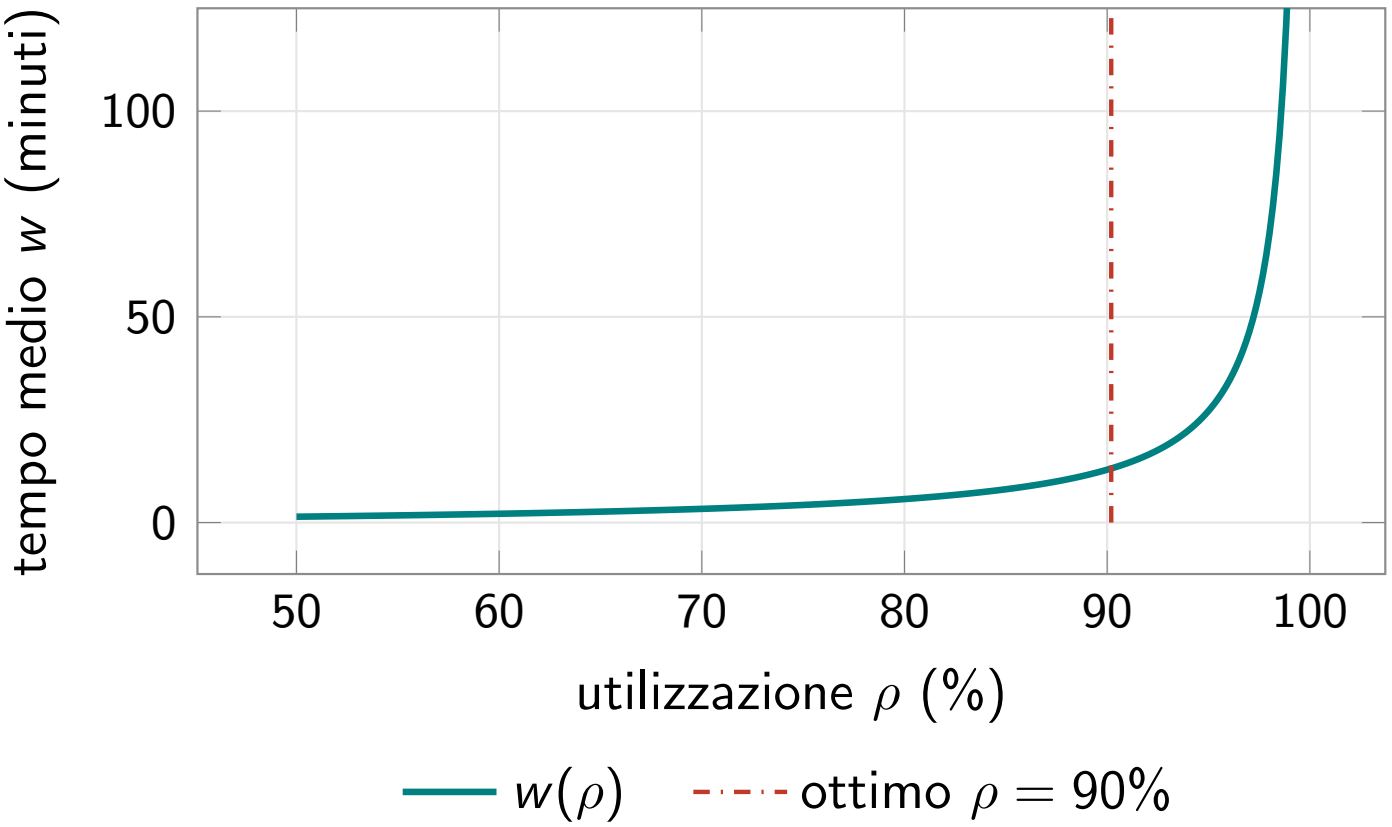
Il costo totale è convesso in μ



Lettura del grafico

- Il costo di capacità $c\mu$ è crescente
- Il costo di attesa $h\lambda/(\mu - \lambda)$ è decrescente
- La loro somma, il costo totale, è convessa
- Il minimo è in $\mu^* = 46{,}6$

Il muro dell'utilizzazione



Oltre il 90% l'attesa esplode in modo iperbolico.

Insight manageriale

- Il muro dell'utilizzazione è forse il grafico più importante da mostrare a un manager
- Tra $\rho = 90\%$ e $\rho = 99\%$ il tempo di attesa si moltiplica per dieci
- "Saturare le risorse" e "dare un buon servizio" sono obiettivi in conflitto **matematico**, non organizzativo
- Nessuna riorganizzazione elimina l'iperbole $w = 1/(\mu - \lambda)$

Riepilogo: Caso di studio e risultati

- Ottimo: $\mu^* = 46{,}58$, costo 153,50 €/h; Gurobi conferma
- All'ottimo $\rho = 90{,}2\%$ e $w = 13{,}1$ minuti
- Vicino al 100% di utilizzazione l'attesa esplode: saturare le risorse e servire bene sono in conflitto

Analisi di sensitività: il prezzo di una promessa

Introduzione: il prezzo di una promessa

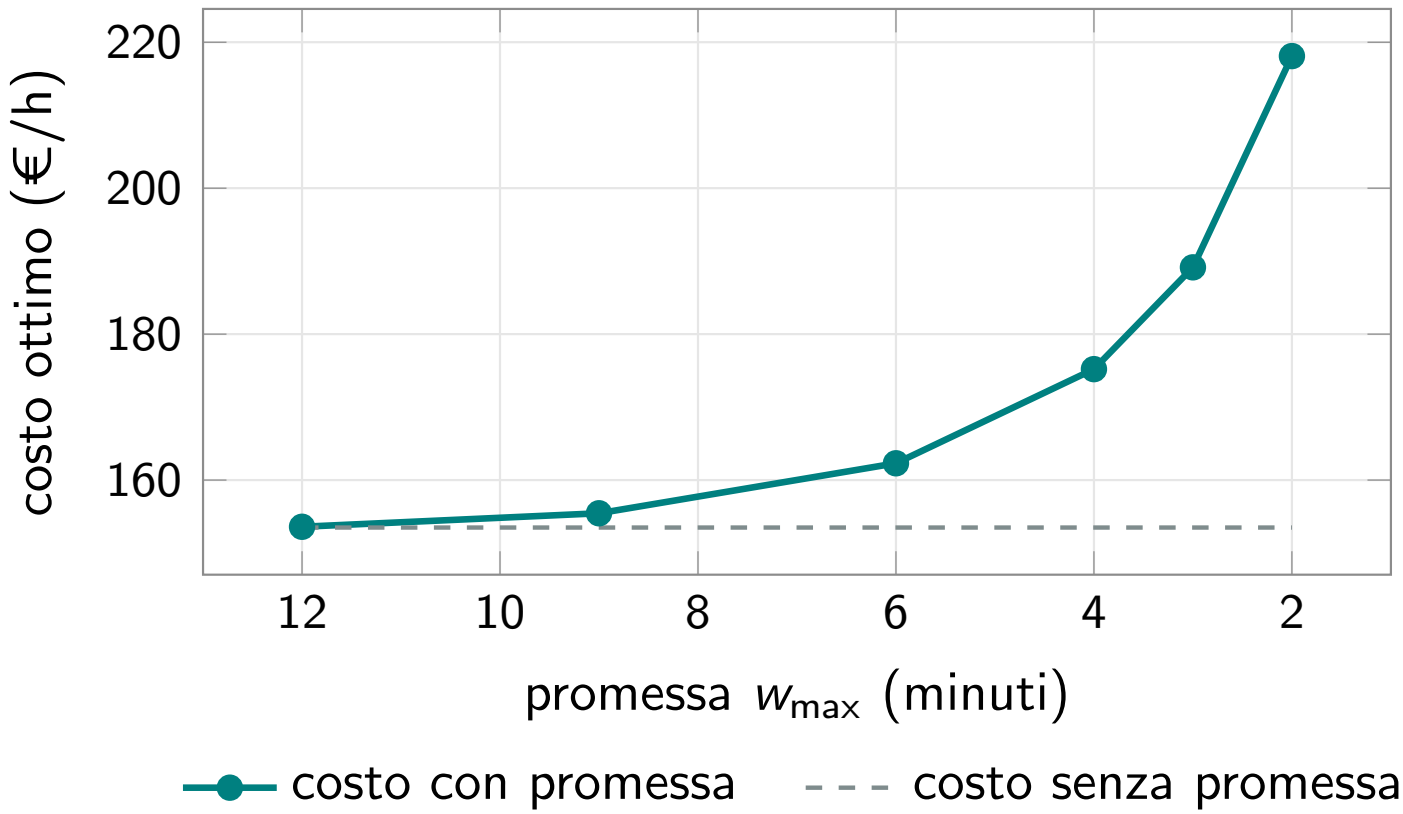
- Il marketing promette un tempo medio sotto una soglia $w_{\max}$
- Quanto costa mantenere questa promessa?
- Calcoliamo il costo marginale della promessa (prezzo ombra)
- Infine proteggiamo il dimensionamento dall'incertezza su λ (robustezza)

La promessa di servizio

- Per avere "tempo medio sotto $w_{\max}$" serve $\mu \geq \lambda + 1/w_{\max}$
- Il vincolo diventa attivo quando è più stringente dell'ottimo economico (13,1 minuti)

```
w_max = 12 min: mu = 47,00  costo 153,60   (quasi gratis)
w_max =  9 min: mu = 48,67  costo 155,45
w_max =  6 min: mu = 52,00  costo 162,30
w_max =  4 min: mu = 57,00  costo 175,20
w_max =  3 min: mu = 62,00  costo 189,15
w_max =  2 min: mu = 72,00  costo 218,10   (+42% sul costo base)
```

Quanto costa una promessa di servizio più ambiziosa



Lettura del grafico

- L'asse è invertito: verso destra la promessa diventa più ambiziosa
- La linea tratteggiata è il costo senza promessa
- Fino a 12 minuti la promessa è quasi gratuita
- Sotto i 4 minuti ogni minuto in meno costa in modo esplosivo

Il costo marginale della promessa (1/2)

- Il prezzo ombra del vincolo cresce in modo esplosivo:
  - da 12 a 9 minuti costa 1,85 €/h
  - da 3 a 2 minuti costa 28,95 €/h
- Quando il vincolo è attivo, $\mu = \lambda + 1/w_{\max}$ e

$$C = c\lambda + c/w_{\max} + h\lambda\, w_{\max}$$

- Da cui (con $w_{\max}$ in ore):

$$\frac{dC}{dw_{\max}} = -\frac{c}{w_{\max}^2} + h\lambda$$

Il costo marginale della promessa (2/2)

- A $w_{\max} = 6$ minuti vale

$$-300 + 63 = -237 \text{ €/h per ora di promessa}$$

  esattamente la stima per differenze finite dello script
- È il numero da dare al marketing *prima* che firmi lo SLA
- SLA = *Service Level Agreement*, l'impegno contrattuale sul livello di servizio

Robustezza

- Se λ è incerto in $[36, 48]$, dimensionare sull'ottimo nominale $\mu^* = 46{,}58$ è **instabile**
- Con $\lambda = 48 > \mu^*$ la coda diverge (costo infinito)
- Minimizzare il costo del caso peggiore dà $\mu_{\text{rob}} = 52{,}90$, con costo worst-case 173,39 €/h
- Si paga un premio assicurativo di $\approx 13\%$ per essere protetti da un disastro
- Le decisioni di capacità vanno sempre verificate contro il *massimo* plausibile della domanda, non solo contro la media

Riepilogo: il prezzo di una promessa

- Una promessa $w \leq w_{\max}$ impone $\mu \geq \lambda + 1/w_{\max}$
- Il costo marginale $dC/dw_{\max} = -c/w_{\max}^2 + h\lambda$ esplode per promesse ambiziose
- Dimensionare sulla domanda media può rendere la coda instabile
- La soluzione robusta costa circa il 13% in più e protegge dal disastro

Esercizi

Introduzione: Esercizi

- Quattro esercizi per consolidare i concetti del capitolo
- Verifica del costo marginale, penalità quadratica, pooling, frontiera capacità-servizio
- Si risolvono con le formule viste e con lo script del laboratorio

Esercizio: verifica

- Ricavare per conto proprio la formula

$$\frac{dC}{dw_{\max}} = -\frac{c}{w_{\max}^2} + h\lambda$$

- Verificarla numericamente a $w_{\max} = 4$ e 9 minuti con differenze finite

Esercizio: penalità quadratica

- Sostituire il costo d'attesa lineare con

$$h \left[\frac{1}{\mu - \lambda} \right]^2$$

- Ricavare μ^* analiticamente
- Confrontare l'utilizzazione ottima con il caso lineare

Esercizio: due code contro una

- Due sportelli separati ricevono $\lambda = 21$ ciascuno
- In alternativa, un unico sistema riceve $\lambda = 42$ con capacità aggregata
- A parità di costo di capacità, quale configurazione dà attese minori?
- È l'argomento quantitativo per il **pooling**

Esercizio: frontiera capacità-servizio

- Tracciare la frontiera (costo, w) variando $w_{\max}$ con continuità
- Individuare il punto oltre il quale ogni minuto promesso in meno costa più di 10 €/h

Riepilogo (1/2)

Modello:

- Coda M/M/1, capacità μ come variabile, vincolo di stabilità $\mu > \lambda$
- Costo totale $c\mu + h\lambda/(\mu - \lambda)$, convesso
- Ottimo $\mu^* = \lambda + \sqrt{h\lambda/c}$: domanda + scorta di sicurezza

Messaggio manageriale:

- L'utilizzazione ottima non è il 100%
- Vicino alla saturazione l'attesa esplode (muro dell'utilizzazione)

Riepilogo (2/2)

Promesse di servizio:

- $w \leq w_{\max}$ diventa $\mu \geq \lambda + 1/w_{\max}$
- Il prezzo ombra esplode per promesse ambiziose: calcolarlo prima di firmare lo SLA

Robustezza:

- Dimensionare sulla domanda massima plausibile, non solo sulla media
- Il premio assicurativo ($\approx 13\%$) evita la divergenza della coda