Simultaneous Routing and Resource Allocation via Dual Decomposition

Lin Xiao Mikael Johansson Stephen Boyd

Information Systems Laboratory
Stanford University

Large-Scale Engineering Networks:
Robustness, Verifiability, and Convergence

IPAM, April 18, 2002

- case study: wireless communication network
  - communication network with nodes connected by wireless links
  - multiple flows, from source to destination nodes
  - total traffic on each link limited by link capacity
  - link capacity is function of communication resource variables such as power, bandwidth, which are limited

  goal: find optimal operation of network, $i.e.$, do *simultaneous routing and resource allocation* (SRRA)

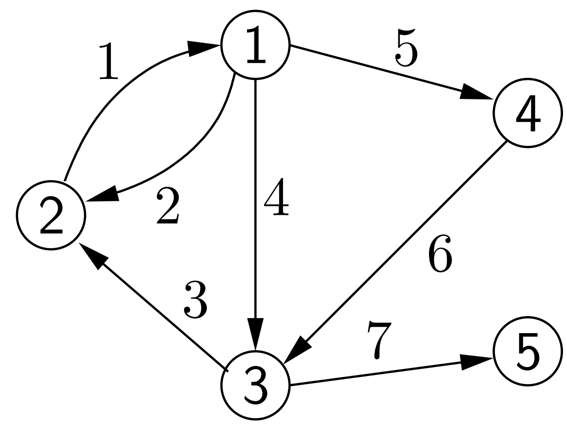
- basic idea: **exploit problem structure via duality**
  - vertical decomposition (dualize coupling constraints between layers)
  - horizontal decomposition (dualize local constraints among neighbors)

Outline

- the *simultaneous routing and resource allocation* (SRRA) problem
  - network flow/routing
  - communication resource allocation
  - formulation of SRRA
  - examples
- solution via dual decomposition (vertical decomposition)
  - formulation of the dual problem
  - subgradient method
  - analytic center cutting-plane method (ACCPM)
- distributed algorithms for subproblems (horizontal decomposition)
  - flow routing
  - resource allocation

Network topology

- directed graph with nodes
  $\mathcal{N} = \{1, \ldots, n\}$, links $\mathcal{L} = \{1, \ldots, m\}$



- $\mathcal{O}(i)$: set of outgoing links at node i
  $\mathcal{I}(i)$: set of incoming links at node i

- incidence matrix $A \in \mathbf{R}^{n \times m}$

$$a_{ik} = \begin{cases} 1, & \text{if } k \in \mathcal{O}(i) \\ -1, & \text{if } k \in \mathcal{I}(i) \\ 0, & \text{otherwise} \end{cases}$$

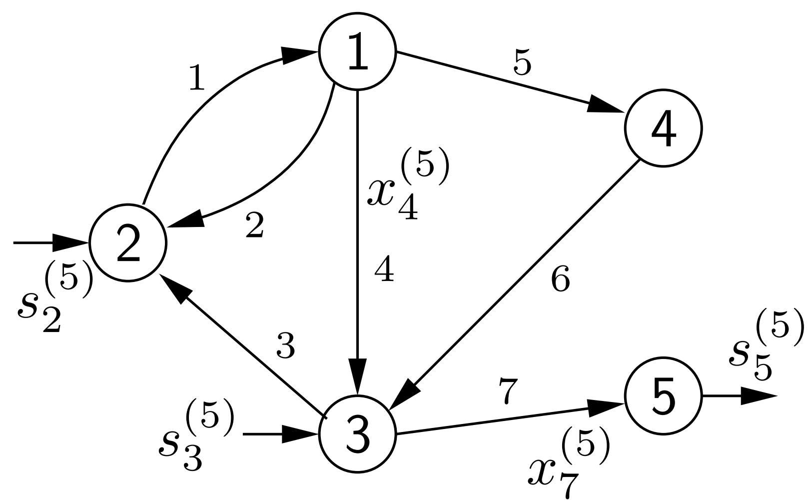
	1	2	3	4	5	6	7
1	-1	1	0	1	1	0	0
2	1	-1	-1	0	0	0	0
3	0	0	1	-1	0	-1	1
4	0	0	0	0	-1	1	0
5	0	0	0	0	0	0	-1

Network flow model

- multiple source/destination pairs
- identify flows by destinations $d \in \mathcal{D} \subseteq \mathcal{N}$
  - $s^{(d)} \in \mathbf{R}^n$: $s_i^{(d)}$ flow from node i to node d
  - $x^{(d)} \in \mathbf{R}^m$: $x_k^{(d)}$ flow on link k, to node d
- flow conservation laws

$$\sum_{k \in \mathcal{O}(i)} x_k^{(d)} - \sum_{k \in \mathcal{I}(i)} x_k^{(d)} = s_i^{(d)}$$

or $\quad Ax^{(d)} = s^{(d)}$



Multicommodity network flow problem

- network flow constraints

$$
\begin{array}{ll}
Ax^{(d)} = s^{(d)}, & \text{flow conservation law} \\
x^{(d)} \succeq 0, & \text{nonnegative flows} \\
t_k = \sum_{d \in \mathcal{D}} x_k^{(d)}, & \text{total traffic on link } k \\
t_k \leq c_k, & \text{capacity constraints}
\end{array}
$$

- one traditional optimal routing problem: with s, c fixed, minimize convex separable function of t, *e.g.*, average or total delay

$$
\text{minimize} \qquad D_{\text{tot}} = \sum_k \frac{t_k}{c_k - t_k}
$$

- another traditional formulation: with c fixed, maximize sum of concave utility functions over source flows:

$$\text{maximize} \qquad U_{\text{tot}} = \sum_{d} \sum_{i \neq d} U_i^{(d)}(s_i^{(d)})$$

- optimization based congestion control (Kelly et al, Low et al, ...)

$$\begin{array}{ll} \text{maximize} & \sum_{r \in \mathcal{R}} U_r(s_r) \\ \text{subject to} & \sum_{r \in \mathcal{S}(l)} s_r \leq c_l, \quad , l \in \mathcal{L} \end{array}$$

  - adjust s_r with fixed routing table; only have capacity constraints
  - TCP running at a faster time scale than IP

- many solution methods, including distributed algorithms by duality (will come back to this later)

Communications model and assumptions

now we consider effect of communication resources ($e.g.$, power, bandwidth) on capacity of the links

θ_k: vector of communication resources for link k, $e.g.$, $\theta_k = (P_k, W_k)$

capacity of link k given by $c_k = \phi_k(\theta_k)$, where ϕ_k is concave, increasing

communication resource limits:

$$C\theta \preceq b, \qquad \theta \succeq 0$$

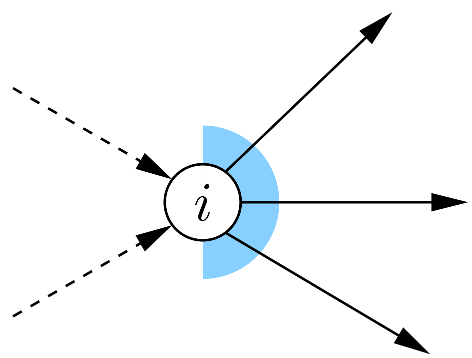
$e.g.$, limits on total transmit power at node, total bandwidth over groups of nodes

Example: Gaussian broadcast channel with FDMA

- communications variables $\theta_k = (P_k, W_k)$, $P_k,\ W_k \geq 0$
- $c_k = \phi_k(P_k, W_k) = W_k \log_2(1 + \frac{P_k}{N_k W_k})$
- total power and bandwidth constraints on each outgoing link:

$$\sum_{k \in \mathcal{O}(i)} P_k \leq P_{\mathrm{tot}}^{(i)}$$

$$\sum_{k \in \mathcal{O}(i)} W_k \leq W_{\mathrm{tot}}^{(i)}$$



Communication resource allocation problem

maximize weighted sum of capacities, subject to resource limits

$$\begin{array}{ll} \text{maximize} & \sum_k w_k c_k = \sum_k w_k \phi_k(\theta_k) \\ \text{subject to} & C\theta \preceq b, \quad \theta \succeq 0 \end{array}$$

- convex problem
- special methods for particular cases, $e.g.$, waterfilling for variable powers, fixed bandwidth

$$\begin{array}{ll} \text{maximize} & \sum_k w_k c_k = \sum_k w_k \phi_k(P_k) \\ \text{subject to} & \sum_k P_k \leq P_{\text{total}}, \quad P_k \geq 0 \end{array}$$

Simultaneous routing and resource allocation

separable convex objective function $f_{\mathrm{net}}(x, s, t) + f_{\mathrm{comm}}(\theta)$

$$
\begin{array}{lll}
\text{minimize} & f_{\mathrm{net}}(x, s, t) + f_{\mathrm{comm}}(\theta) & \\
\text{subject to} & Ax^{(d)} = s^{(d)}, & \text{flow conservation} \\
& x^{(d)} \succeq 0, & \text{nonnegative flows} \\
& t_k = \sum_{d \in \mathcal{D}} x_k^{(d)}, & \text{total traffic on links} \\
& t_k \leq \phi_k(\theta_k), & \text{capacity constraints} \\
& C\theta \preceq b, \quad \theta \succeq 0 & \text{resource limits}
\end{array}
$$

- a **convex optimization problem** with variables $x,\ s,\ t,\ \theta$
- when communication resource allocation θ is fixed, get convex multicommodity flow problem

Examples

Minimum total power/bandwidth SRRA:

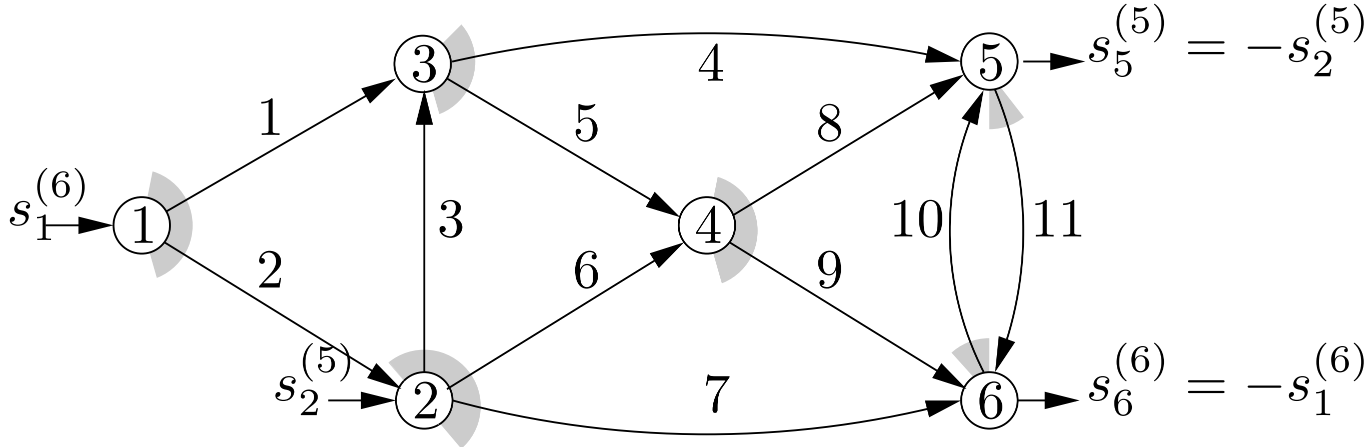
- source-sink vectors $s^{(d)}$ given
- SRRA objective function: $w^T\theta$, $w_i = \begin{cases} 1 & \theta_i \text{ is a power variable,} \\ 0 & \text{otherwise} \end{cases}$

variation: minimum total required bandwidth

Maximum utility SRRA:

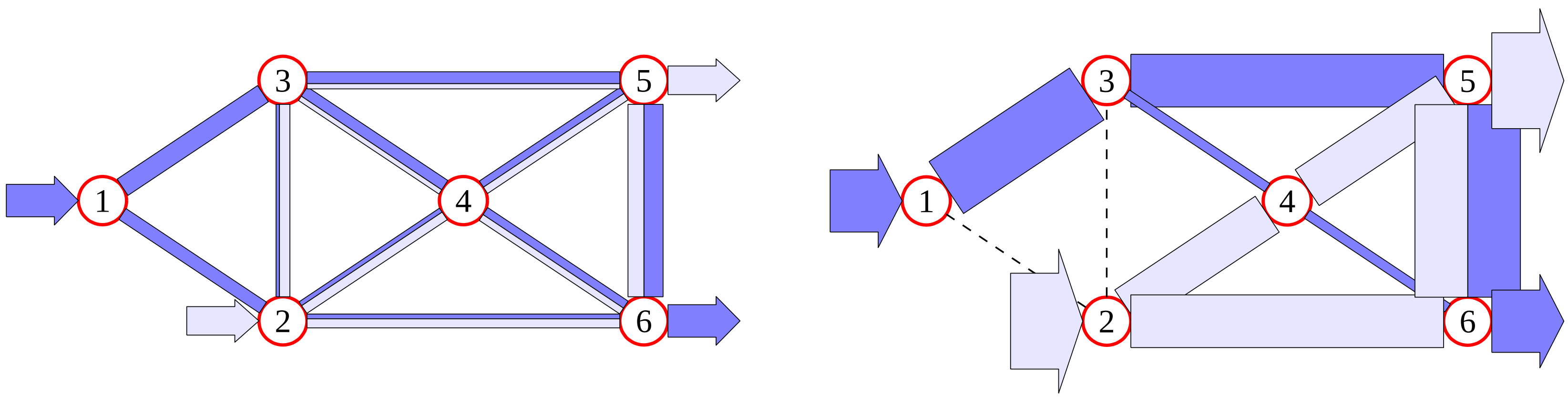
- total utility given by $U(s) = \sum_d \sum_{i \neq d} U_i^{(d)}(s_i^{(d)})$

An example with FDMA



- total transmit power at each node: $P_{\text{tot}}^{(i)} = 1$
- total bandwidth, over all links in network: $W_{\text{tot}} = 11$
- receiver noise spectral densities: $N_k = 0.1$
- objective: maximize sum of flows: $s_1^{(6)} + s_2^{(5)}$

Optimal routing & resource allocation



- left: allocate power and bandwidth evenly across links, then optimize flow; get $s_1^{(6)} + s_2^{(5)} = 1.27$

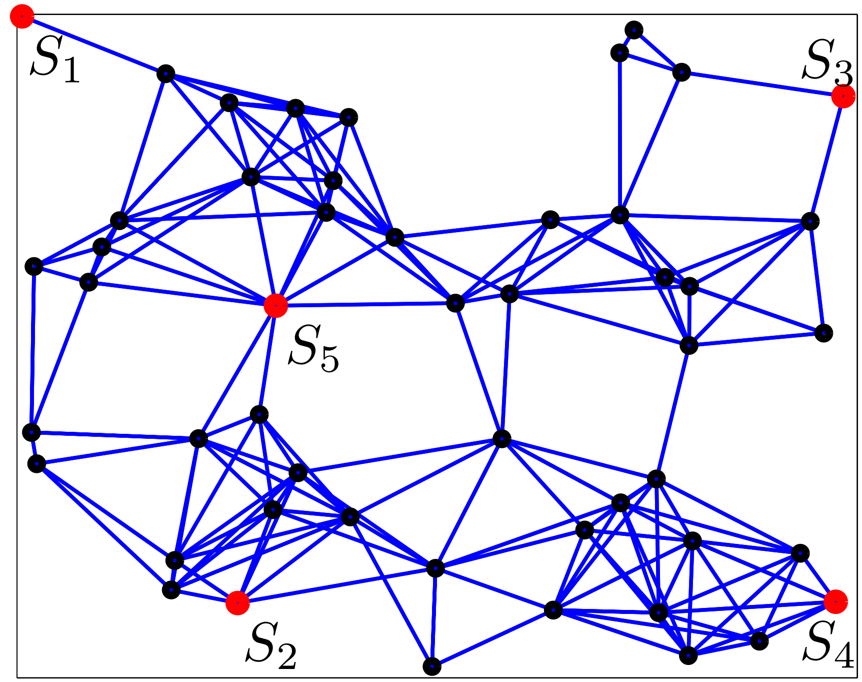
- right: solve SRRA problem (46 variables); get $s_1^{(6)} + s_2^{(5)} = 8.22$

SRRA gives significant performance improvement, sparse optimal routes (load/utility dependent topology: choose an effienent subgraph)

Solution methods

- real-world problems: hundreds of nodes, thousands of links
- general methods for convex problems: interior point methods
- can exploit structure in problem:
  - A, and often C, are very sparse
  - most constraints are local
- for real-world implementation: distributed algorithms

A larger example



- 50 nodes, 340 links
- 5 destination nodes, 20 source/destination pairs
- 2060 variables (1720 flow variables, 340 power variables)

- generate random network topology
  - nodes uniformly distributed on a square
  - two nodes communicate if distance smaller than threshold
  - randomly choose source and destination nodes

- bandwidth allocation fixed; only allocate transmit power p_k

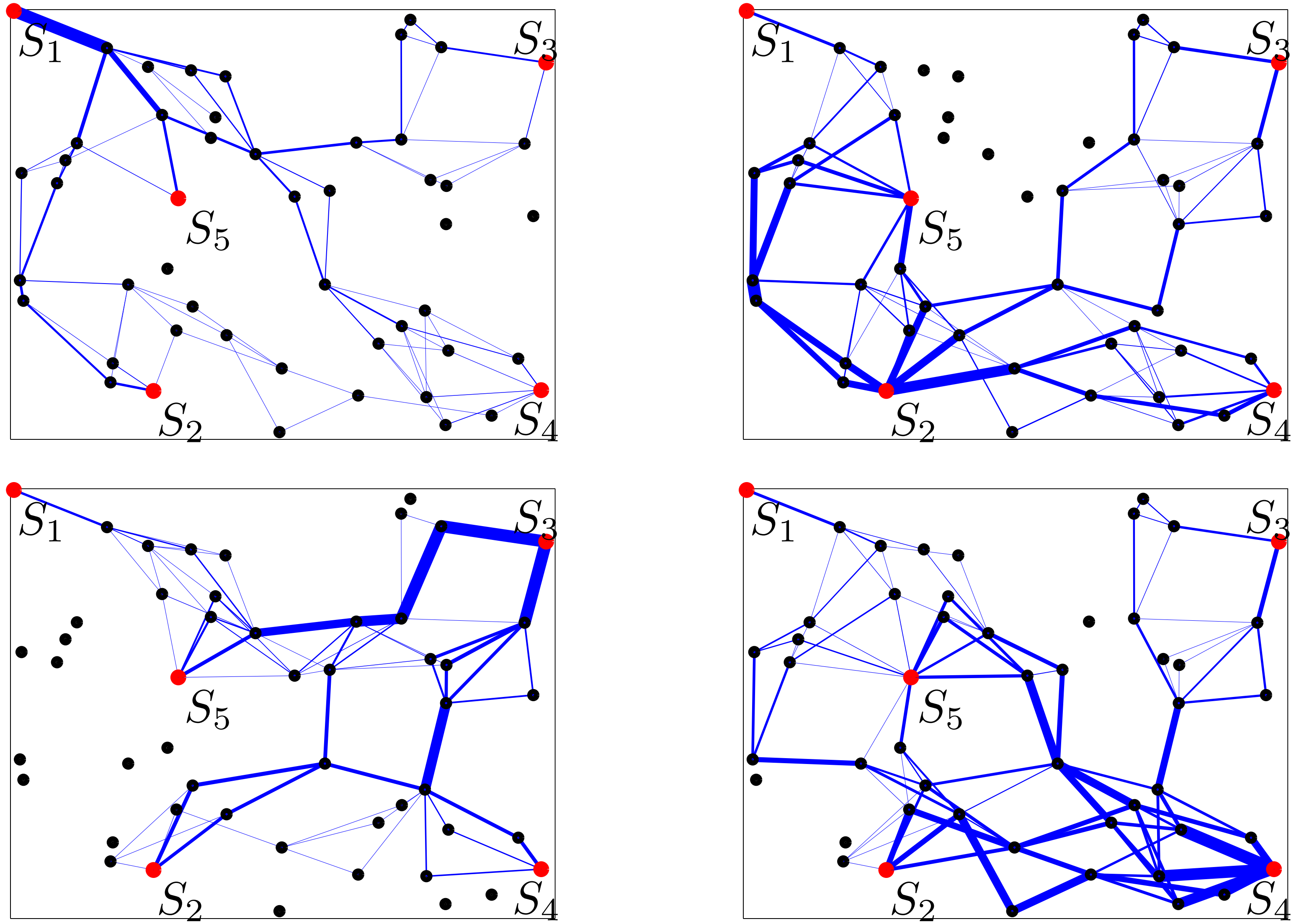
- total power limit at each node $\displaystyle\sum_{k \in \mathcal{O}(i)} p_k \leq p_{\mathrm{tot}}^i$

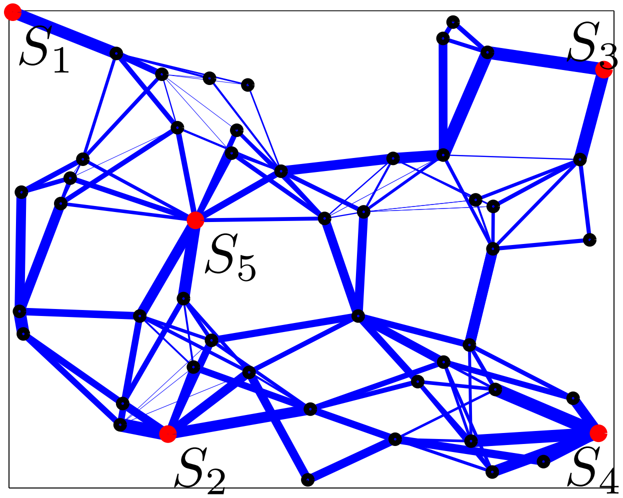
- power path loss model $P_k = p_k K \left(\dfrac{d_0}{d_k}\right)^2$

- noise power N_i uniformly distributed on $[\underline{N}, \overline{N}]$

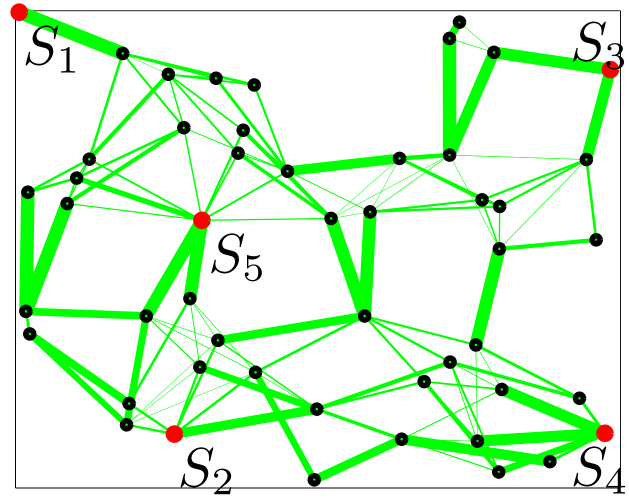
- source utility function $U(s) = \displaystyle\sum_d \sum_{i \neq d} \log s_i^{(d)}$

Optimal routes

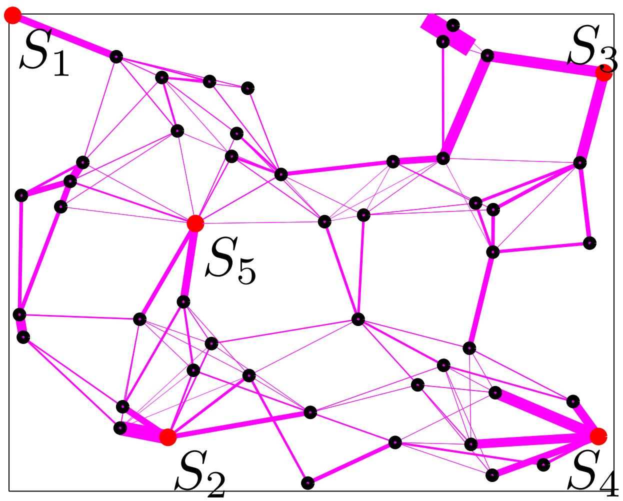




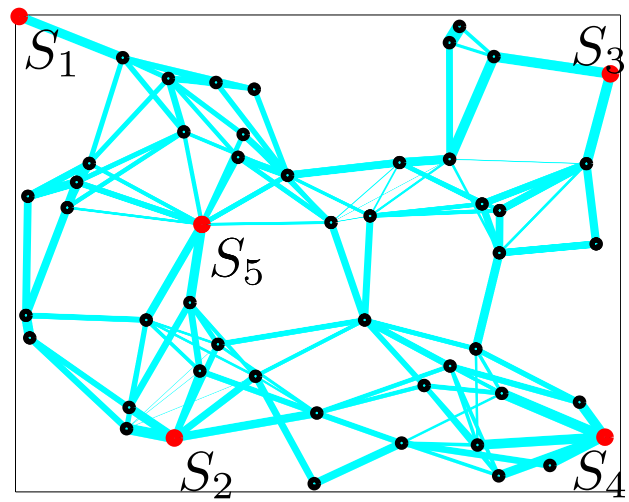
aggregate flow



power allocation



SNRs



link capacities

Comparison with uniform power allocation

i	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
1	-2.26	1.03	0.88	1.01	1.37
2	0.56	-13.95	1.73	9.59	5.92
3	0.54	2.07	-6.61	1.97	4.14
4	0.54	6.70	1.55	-16.34	4.20
5	0.62	4.15	2.45	3.77	-15.63

Table 1: Source-sink flows $s_i^{(d)}$ with fixed capacity routing (uniform power allocation), total utility: 12.77

i	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
1	-3.88	1.11	0.92	1.12	1.13
2	1.03	-16.05	2.93	6.98	6.97
3	0.84	2.69	-9.43	2.69	2.77
4	0.96	4.80	2.46	-18.23	4.80
5	1.05	7.45	3.12	7.44	-15.67

Table 2: Source-sink flows $s_i^{(d)}$ with simultaneous routing and resource allocation, total utility: 17.27

Outline

- the *simultaneous routing and resource allocation* (SRRA) problem
  - network flow/routing
  - communication resource allocation
  - formulation of SRRA
  - examples
- **solution via dual decomposition (vertical decomposition)**
  - formulation of the dual problem
  - subgradient method
  - analytic center cutting-plane method (ACCPM)
- distributed algorithms for subproblems (horizontal decomposition)
  - flow routing
  - resource allocation

Exploiting structure via dual decomposition

structure of SRRA problem

- objective separable in network flow and communications variables
- only capacity constraints couple $x,\ s,\ t$ and θ

dual decomposition (Lagrange relaxation)

- relax coupling capacity constraints by introducing Lagrange multipliers
- decompose SRRA into two subproblems, both highly structured, efficient algorithms exist for each (dual decomposition again)
- subproblems coordinated by master dual problem

The SRRA problem

$$
\begin{array}{lll}
\text{minimize} & f_{\text{net}}(x, s, t) + f_{\text{comm}}(\theta) & \\
\text{subject to} & Ax^{(d)} = s^{(d)}, & \text{flow conservation} \\
 & x^{(d)} \succeq 0, & \text{nonnegative flows} \\
 & t_k = \sum_{d \in \mathcal{D}} x_k^{(d)}, & \text{total traffic on links} \\
 & t_k \leq \phi_k(\theta_k), & \textbf{capacity constraints} \\
 & C\theta \preceq b, \quad \theta \succeq 0 & \text{resource limits}
\end{array}
$$

Dual decomposition

- introduce multiplier $\lambda \in \mathbf{R}_+^m$ only for coupling constraints

$$\begin{aligned} L(x,s,t,\theta,\lambda) &= f_{\rm net}(x,s,t) + f_{\rm comm}(\theta) + \lambda^T(t-\phi(\theta)) \\ &= \Big(f_{\rm net}(x,s,t) + \lambda^T t\Big) + \Big(f_{\rm comm}(\theta) - \lambda^T\phi(\theta)\Big), \end{aligned}$$

- dual function

$$\begin{aligned} g(\lambda) &= \inf\left\{ L(x,s,t,\theta,\lambda) \,\middle|\, \begin{array}{l} Ax^{(d)} = s^{(d)},\ x^{(d)} \succeq 0,\ \sum_{d\in\mathcal{D}} x^{(d)} = t \\ C\theta \preceq b,\ \theta \succeq 0 \end{array} \right\} \\ &= g_{\rm net}(\lambda) + g_{\rm comm}(\lambda) \end{aligned}$$

$$\begin{aligned} g_{\rm net}(\lambda) &= \inf\left\{ f_{\rm net}(x,s,t) + \lambda^T t \,\middle|\, Ax^{(d)} = s^{(d)},\ x^{(d)} \succeq 0,\ \sum_{d\in\mathcal{D}} x^{(d)} = t \right\} \\ g_{\rm comm}(\lambda) &= \inf\left\{ f_{\rm comm}(\theta) - \lambda^T\phi(\theta) \,\middle|\, C\theta \preceq b,\ \theta \succeq 0 \right\} \end{aligned}$$

The dual problem SRRA*

- master dual problem (coordinate capacity prices)

$$\begin{array}{ll} \text{maximize} & g(\lambda) = g_{\rm net}(\lambda) + g_{\rm comm}(\lambda) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- network flow subproblem (evaluate $g_{\rm net}(\lambda)$)

$$\begin{array}{ll} \text{minimize} & f_{\rm net}(x,s,t) + \lambda^T t \\ \text{subject to} & Ax^{(d)} = s^{(d)}, \quad x^{(d)} \succeq 0 \\ & t = \sum_{d\in\mathcal{D}} x^{(d)} \end{array}$$

- resource allocation subproblem (evaluate $g_{\rm comm}(\lambda)$)

$$\begin{array}{ll} \text{minimize} & f_{\rm comm}(\theta) - \lambda^T \phi(\theta) \\ \text{subject to} & C\theta \preceq b, \quad \theta \succeq 0 \end{array}$$

economic interpretation

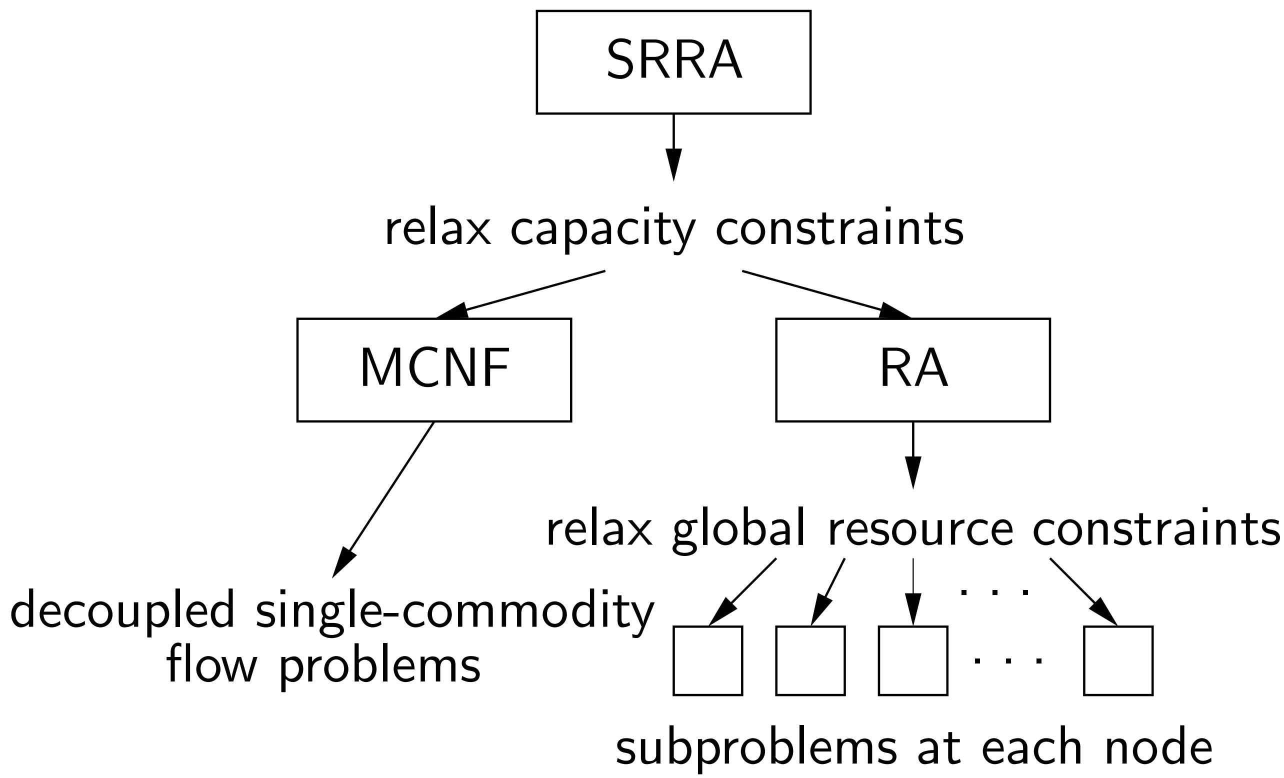
Solving the subproblems

multicommodity flow problem: standard, efficient algorithms exist

resource allocation problem

- structure
  - objective often separable
  - most constraints are local
  - few global constraints, *e.g.*, total bandwidth
- second-level dual decomposition
  - relax global resource constraints
  - subproblems local (at nodes, links)

Hierarchical dual decomposition



subproblems can be solved in parallel, distributed algorithms also exist

Solving SRRA via the dual

- strong duality from constraint qualification
- dual function often nonsmooth (primal objective not strict convex), recovering feasible primal optimal solution is not straightfoward
  - add small regularization terms (strict convex)
  - augmented Lagrangian, proximal bundle method
  - ergodic sequences

Solving SRRA*

non-smooth convex optimization problem, two class of methods

- subgradient (supergradient) methods (Shor, ...)
- cutting plane methods, $e.g.$, ACCPM (Goffin, Vial, Luo, Ye, ...)

all need supergradient information

for SRRA* problem

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

the supergradient $h(\lambda)$ is readily given by $h(\lambda) = t^\star(\lambda) - \phi(\theta^\star(\lambda))$

Subgradient methods

for $k = 1, 2, 3, \ldots$, find supergradient $h^{(k)}$

$$\lambda^{(k+1)} = \left(\lambda^{(k)} + a_k h^{(k)}\right)_+$$

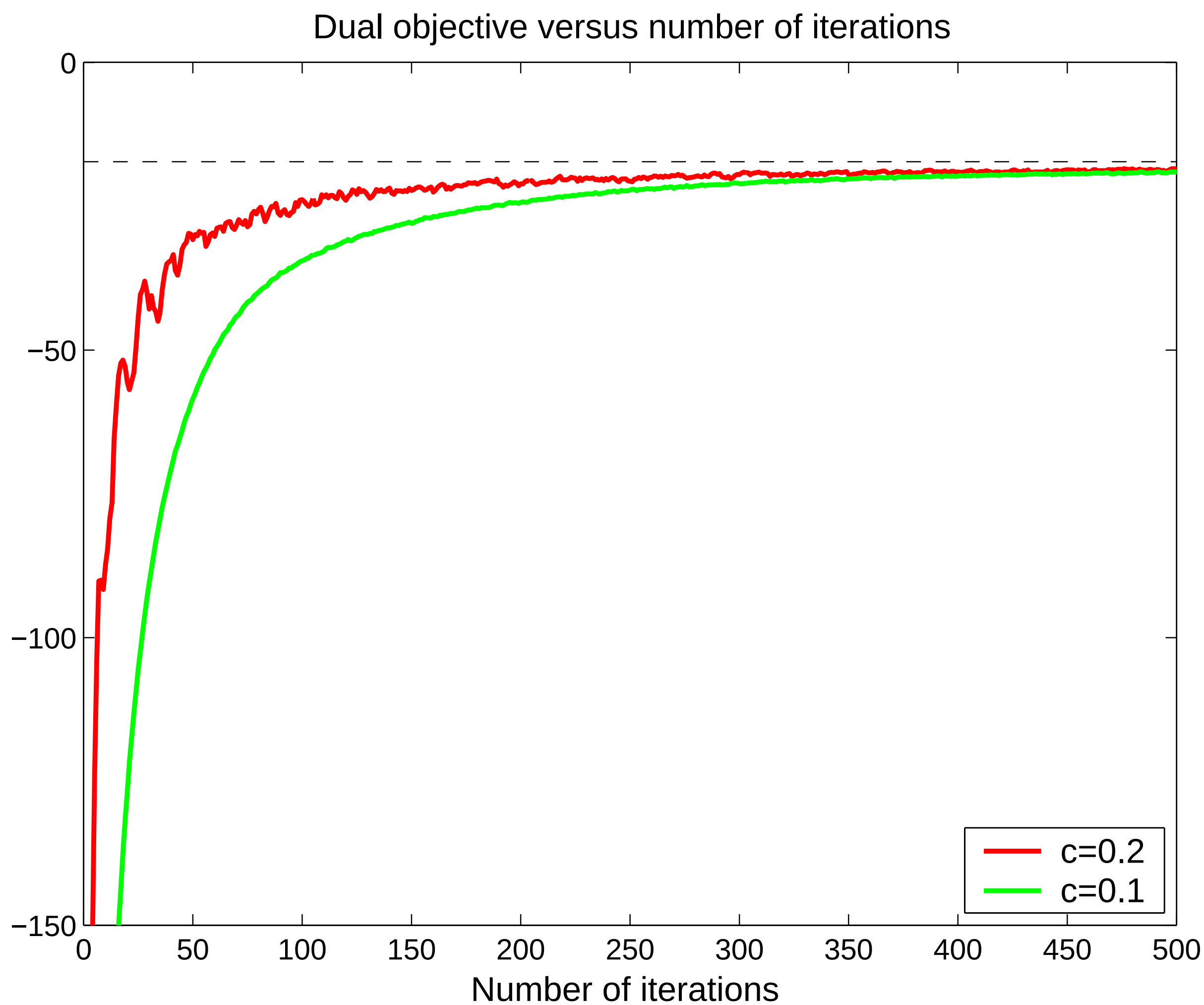
where step size a_k satisfies

$$a_k \geq 0, \qquad a_k \to 0, \qquad \sum_{k=1}^{\infty} a_k = \infty,$$

for example, $a_k = \dfrac{c}{k}$

- update price (dual variable) locally at each link; distributed algorithm

Dual objective versus number of iterations



Analytic center cutting-plane method (ACCPM)

- for $k = 1, 2, 3, \ldots$, compute $g(\lambda^{(k)})$ and supergradient $h^{(k)}$, so

$$g(\lambda) \leq g(\lambda^{(k)}) + {h^{(k)}}^T (\lambda - \lambda^{(k)})$$

each is a linear inequality in the epigraph space $(g(\lambda), \lambda) \in \mathbf{R}^{m+1}$

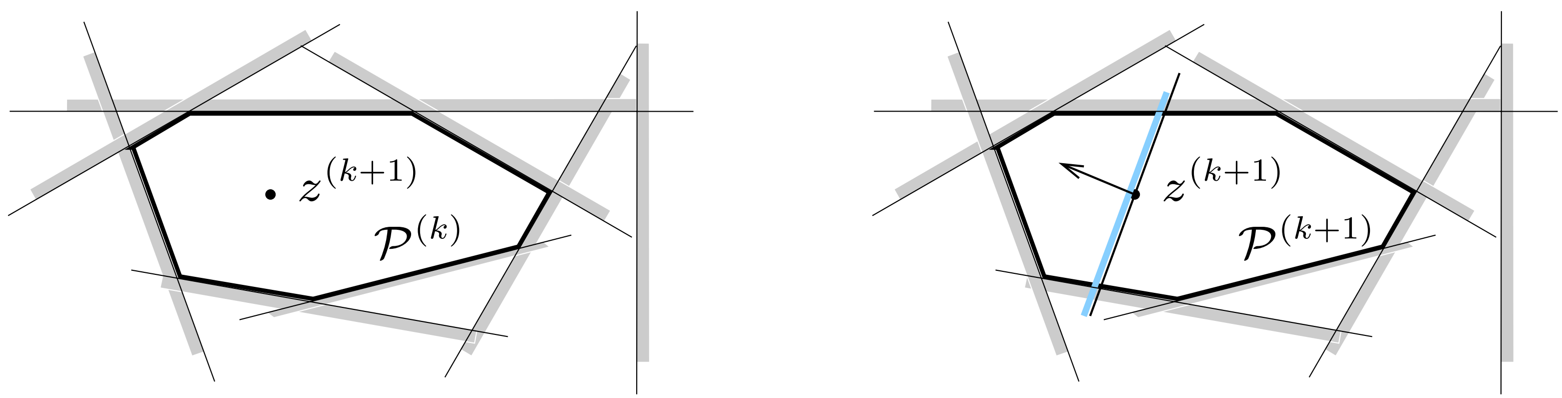
- at step k, they form a polyhedron (the localization set)

$$\mathcal{P}^{(k)} = \left\{ z \;\middle|\; a^{(i)T} z \leq b^{(i)},\ i = 1, \ldots, k,\ z \in \mathbf{R}^{m+1} \right\}$$

the optimal solution $z^\star = (g(\lambda^\star), \lambda^\star)$ lies inside this polyhedron

- compute the analytic center of $\mathcal{P}^{(k)}$

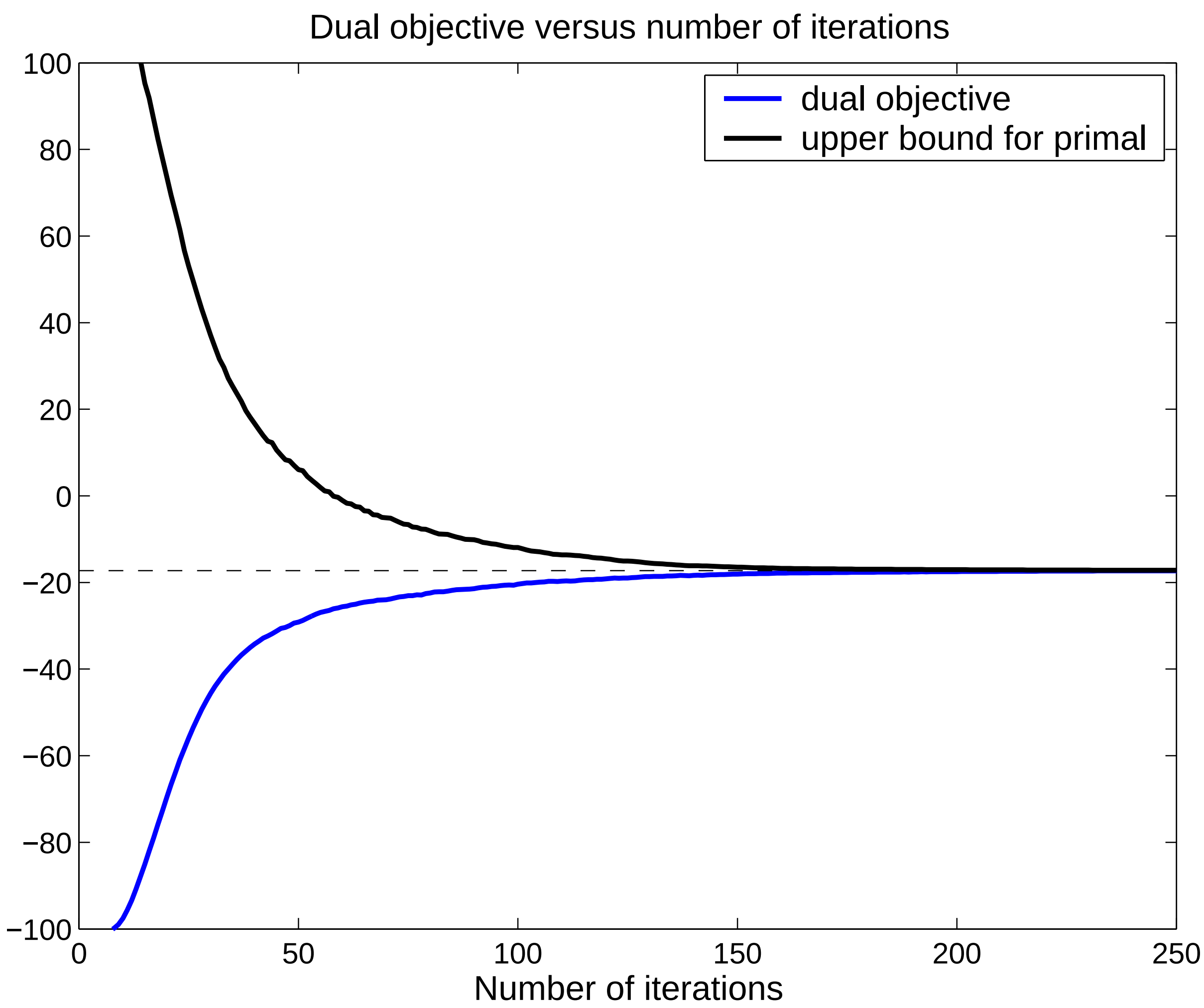
$$z^{(k+1)} = \arg\max_z \sum_{i=1}^{k} \log(b^{(i)} - a^{(i)T} z)$$



- choose $\lambda^{(k+1)}$ as the query point; compute $g(\lambda^{(k+1)})$ and $h^{(k+1)}$

- refine the localization set by adding a halfspace constraint passing through $z^{(k+1)}$ (can have deeper cut)

Dual objective versus number of iterations



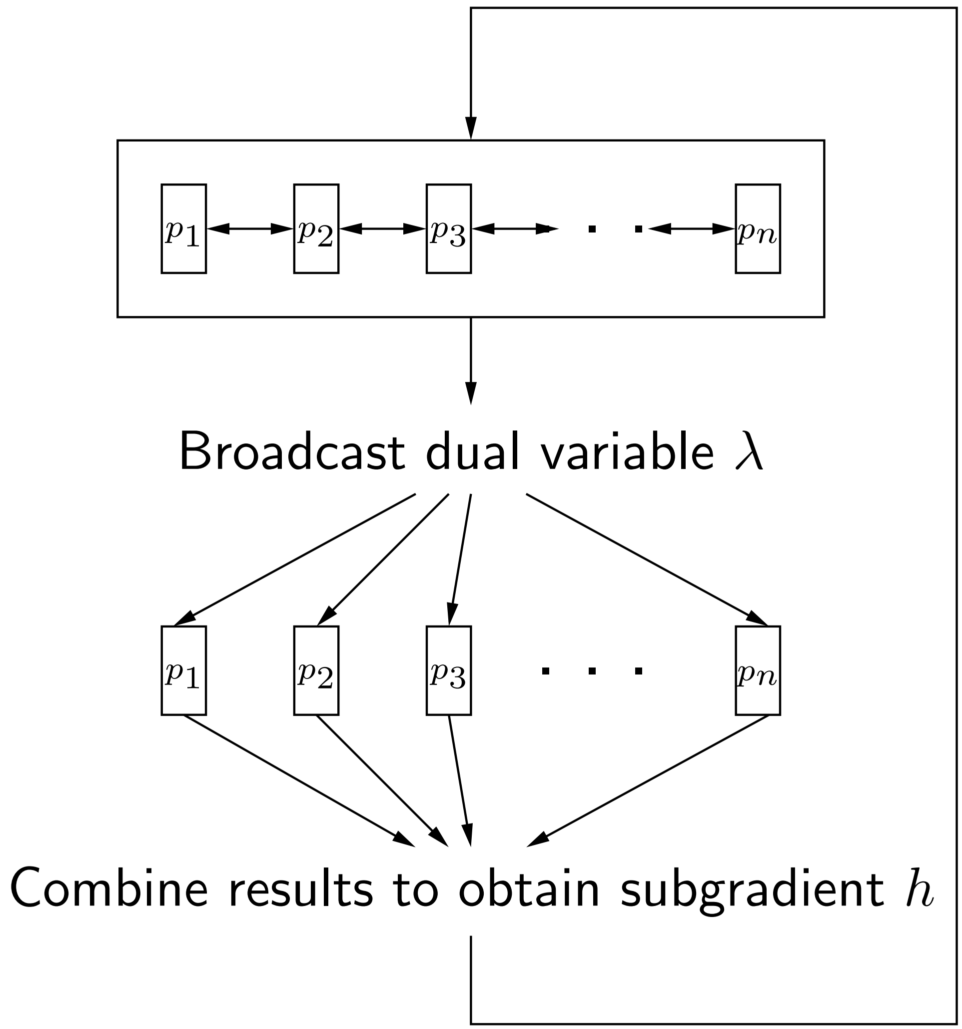
Parallel ACCPM running on multiple processors

Compute AC λ

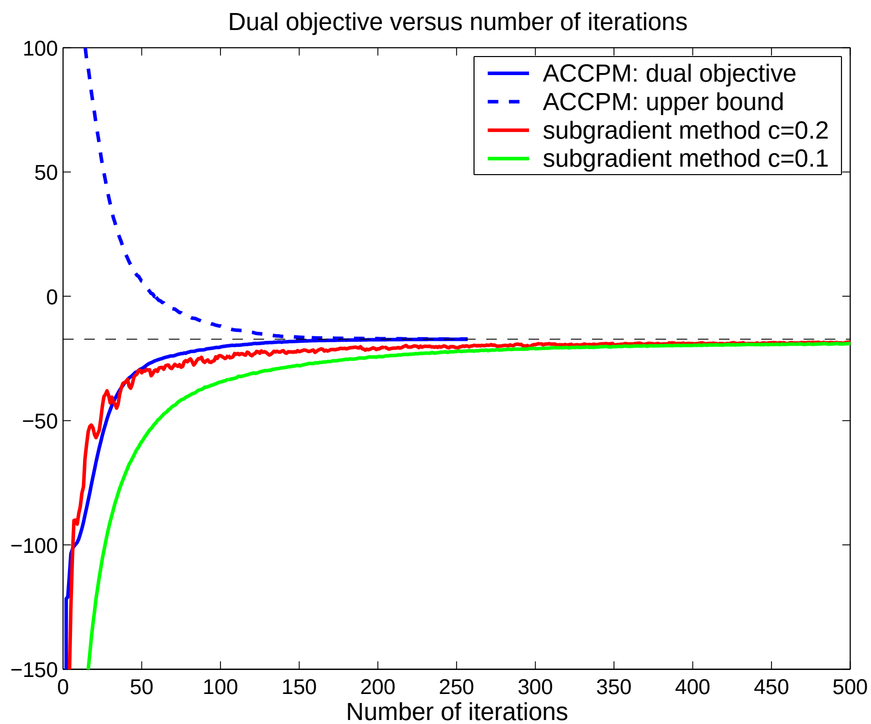
(ScaLAPACK)

Routing and RA

(Sparse solver)



Subgradient methods versus ACCPM



- subgradient methods: slow convergence, but fully distributed
- ACCPM: fast convergence, but needs centralized coordination
- hybrid algorithms possible (??)

Outline

- the *simultaneous routing and resource allocation* (SRRA) problem
  - network flow/routing
  - communication resource allocation
  - formulation of SRRA
  - examples
- solution via dual decomposition (vertical decomposition)
  - formulation of the dual problem
  - subgradient method
  - analytic center cutting-plane method (ACCPM)
- **distributed algorithms for subproblems (horizontal decomposition)**
  - flow routing
  - resource allocation

Distributed routing algorithm

single commodity flow routing problem

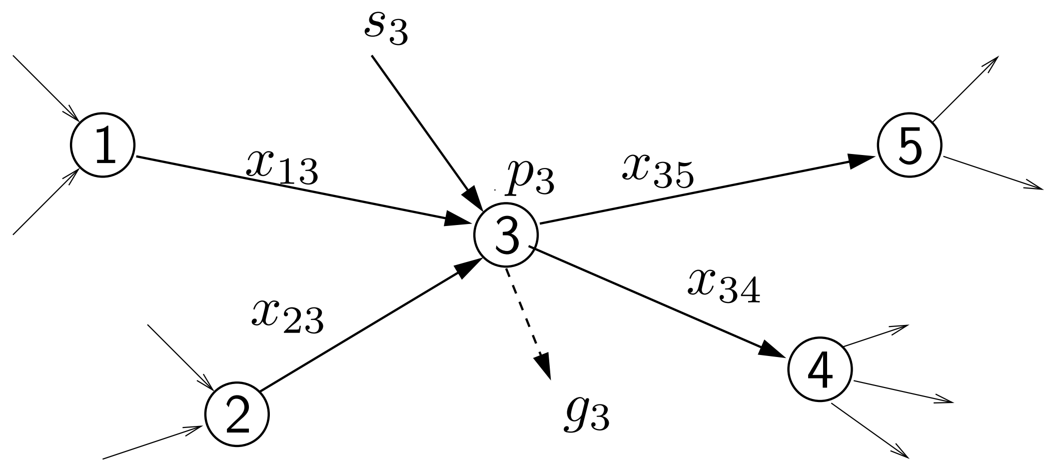
$$\begin{array}{ll} \text{minimize} & \sum_i f(s_i) + \lambda^T x \\ \text{subject to} & Ax = s, \quad x \succeq 0 \end{array}$$

relax flow conservation law at each node by introducing Lagrange multiplier p_i, the dual problem:

$$\text{maximize} \quad q(p)$$

where the dual function

$$q(p) = \inf_{x \succeq 0} \sum_i f(s_i) + \lambda^T x + p^T (Ax - s)$$



subgradient of $q(p)$ in the coordinate of p_3 is given by the "surplus"

$$g_3 = s_3 + x_{13} + x_{23} - x_{34} - x_{35}$$

coordinate ascent: fix other p_i's, adjust p_3 and its incident flow variables to make $g_3 = 0$ (many variations, Bertsekas et al, ...)

- for shortest path problem, exactly the same as distributed Bellman-Ford algorithm: dual variable as cost-to-go value function

- electical circuit analogy: KVL and KCL

Distributed algorithms for resource allocation

consider a simple version

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{n} f_i(x_i) \\ \text{subject to} & x_1 + x_2 + \cdots + x_n \leq T \end{array}$$

where f_i's are **concave** utility function, T is the total resource

- the dual algorithm (pricing)

$$\begin{array}{ll} \text{Let} & x_i(\lambda) = \arg\max_{x_i}(f_i(x_i) - \lambda x_i) \\ \text{update price} & \lambda^+ = \lambda + \alpha \left(\sum_i x_i(\lambda) - T \right) \end{array}$$

- a primal algorithm

shadow price $\lambda_i(x_i) = f_i'(x_i)$

reallocation $x_i^+ = x_i + \alpha \left(\lambda_i(x_i) - \frac{1}{n} \sum_i \lambda_i(x_i) \right)$

- a center-free algorithm (Ho et al, 1980)

shadow price $\lambda_i(x_i) = f_i'(x_i)$

reallocation $$x_i^+ = x_i + \alpha_{i,i-1} \left(\lambda_i(x_i) - \lambda_{i-1}(x_{i-1}) \right) + \alpha_{i,i+1} \left(\lambda_i(x_i) - \lambda_{i+1}(x_{i+1}) \right)$$

variations: communicate shadow prices not only with neighbors

theme: convexity makes all sorts of things possible ...

Some insight on the dual problem

- naturally decoupled simple constraints (**R** or $\mathbf{R}_+$)
- coordinate ascent method for dual problem

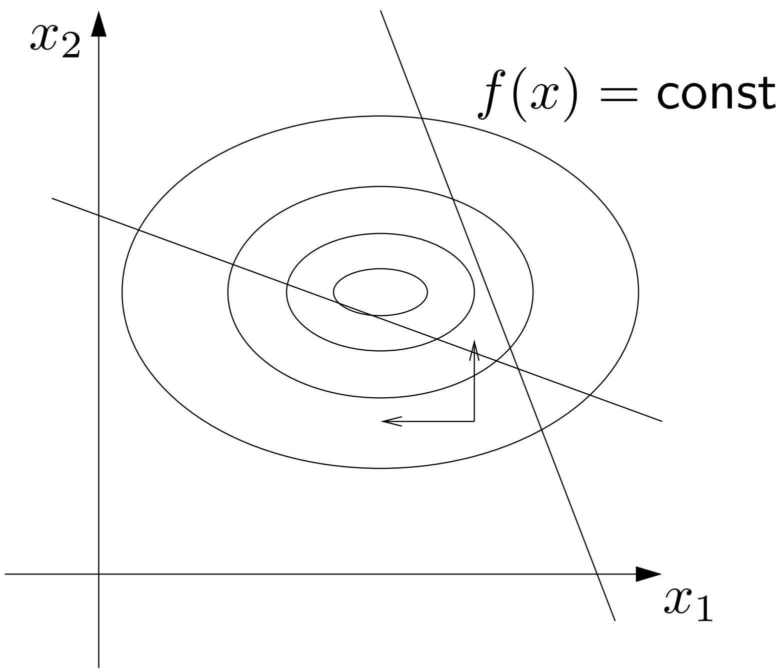
$$\text{maximize} \quad q_i(p_i) = q(p_1, \ldots, p_{i-1}, p_i, p_{i+1}, \ldots, p_m)$$

  with $p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_m$ fixed

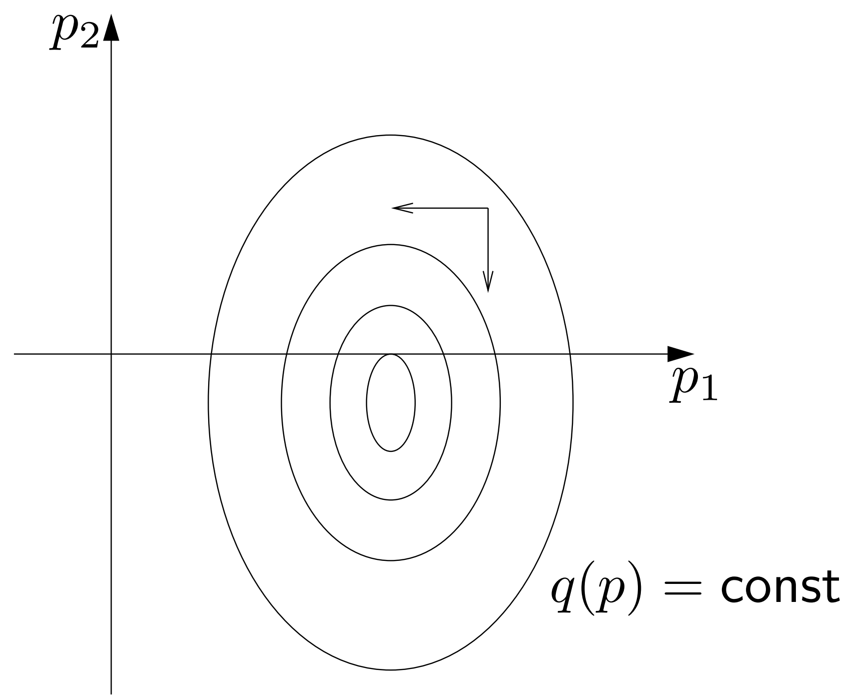
- for networked system, maximization of $q_i(p_i)$ often only need information from a few neighbors, $e.g.$, p_{i-1} and p_{i+1}

suitable for distributed algorithms

Coordinate ascent for primal and dual



primal



dual

Summary

- model and assumptions for wireless data networks
  - capacitated multicommodity flow model
  - capacity constraints concave in communications variables
  - communications resource limits
- SRRA: convex optimization problem
- efficiently solved via dual decomposition; subgradient method, ACCPM
- distributed algorithms for solving subproblems
- extensions
  - asynchronous distributed algorithms
  - dynamic routing and resource allocation

Essential idea

exploit structure of networked system via duality

- vertical decomposition (dualize coupling constraints between layers)
- horizontal decomposition (dualize local constraints among neighbors)

often working at different time scale