Subgradient Methods

- subgradient method and stepsize rules
- convergence results and proof
- projected subgradient method
- projected subgradient method for dual
- optimal network flow

Subgradient method

subgradient method is simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- $x^{(k)}$ is the kth iterate
- $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- $\alpha_k > 0$ is the kth step size

Step size rules

step sizes are fixed ahead of time

- *constant step size:* $\alpha_k = h$ (constant)
- *constant step length:* $\alpha_k = h/\|g^{(k)}\|_2$ (so $\|x^{(k+1)} - x^{(k)}\|_2 = h$)
- *square summable but not summable:* step sizes satisfy

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

- *nonsummable diminishing:* step sizes satisfy

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

Convergence results

- we assume $\|g\|_2 \leq G$ for all $g \in \partial f$ (equivalent to Lipschitz condition on f), $p^\star = \inf_x f(x) > -\infty$
- define $f_{\rm best}^{(k)} = \min_{i=0,\ldots,k} f(x^{(i)})$, best value found in k iterations, and $\bar{f} = \lim_{k\to\infty} f_{\rm best}^{(k)}$

- *constant step size:* $\bar{f} - p^\star \leq G^2 h/2$, $i.e.$, **converges to** $G^2h/2$**-suboptimal** (converges to $p^\star$ if f differentiable, h small enough)
- *constant step length:* $\bar{f} - p^\star \leq Gh/2$, $i.e.$, **converges to** $Gh/2$**-suboptimal**
- *diminishing step size rule:* $\bar{f} = p^\star$, $i.e.$, **converges**

Convergence proof

key quantity: *Euclidean distance to the optimal set*, not the function value

let $x^\star$ be any minimizer of f

$$\begin{aligned}
\|x^{(k+1)} - x^\star\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^\star\|_2^2 \\
&= \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^\star) + \alpha_k^2 \|g^{(k)}\|_2^2 \\
&\le \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k (f(x^{(k)}) - p^\star) + \alpha_k^2 \|g^{(k)}\|_2^2
\end{aligned}$$

using $p^\star = f(x^\star) \ge f(x^{(k)}) + g^{(k)T}(x^\star - x^{(k)})$

applying recursively, and using $\|g^{(i)}\|_2 \leq G$, we get

$$\|x^{(k+1)} - x^\star\|_2^2 \leq \|x^{(1)} - x^\star\|_2^2 - 2\sum_{i=1}^k \alpha_i(f(x^{(i)}) - p^\star) + G^2 \sum_{i=1}^k \alpha_i^2$$

now we use

$$\sum_{i=1}^k \alpha_i(f(x^{(i)}) - p^\star) \geq (f_{\rm best}^{(k)} - p^\star)\left(\sum_{i=1}^k \alpha_i\right)$$

to get

$$f_{\rm best}^{(k)} - p^\star \leq \frac{\|x^{(1)} - x^\star\|_2^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i}.$$

constant step size: for $\alpha_k = h$ we get

$$f_{\rm best}^{(k)} - p^\star \le \frac{\|x^{(1)} - x^\star\|_2^2 + G^2 k h^2}{2kh}$$

righthand side converges to $G^2 h/2$ as $k \to \infty$

square summable but not summable step sizes:
suppose step sizes satisfy

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

then

$$f_{\rm best}^{(k)} - p^\star \leq \frac{\|x^{(1)} - x^\star\|_2^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}$$

as $k \to \infty$, numerator converges to a finite number, denominator converges to ∞, so $f_{\rm best}^{(k)} \to p^\star$

Example: Piecewise linear minimization

minimize $\quad f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$

to find a subgradient of f: find index j for which

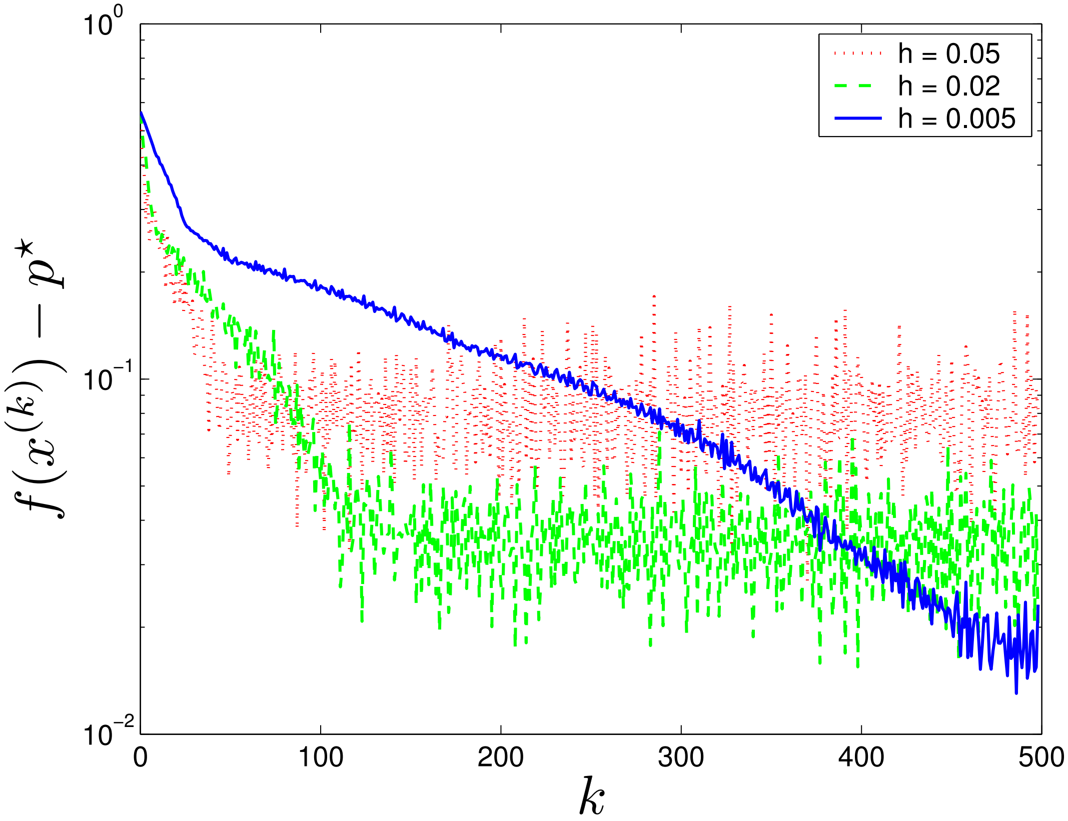
$$a_j^T x + b_j = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

and take $g = a_j$

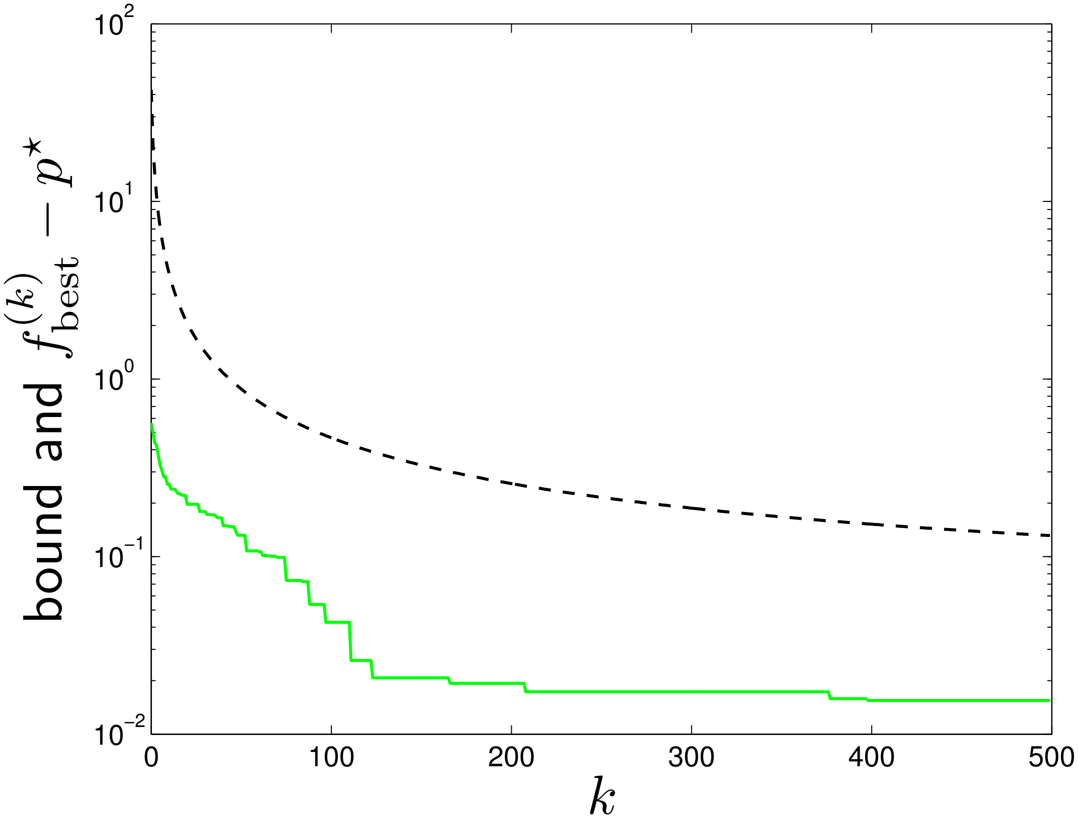
subgradient method: $x^{(k+1)} = x^{(k)} - \alpha_k a_j$

problem instance with $n = 10$ variables, $m = 100$ terms

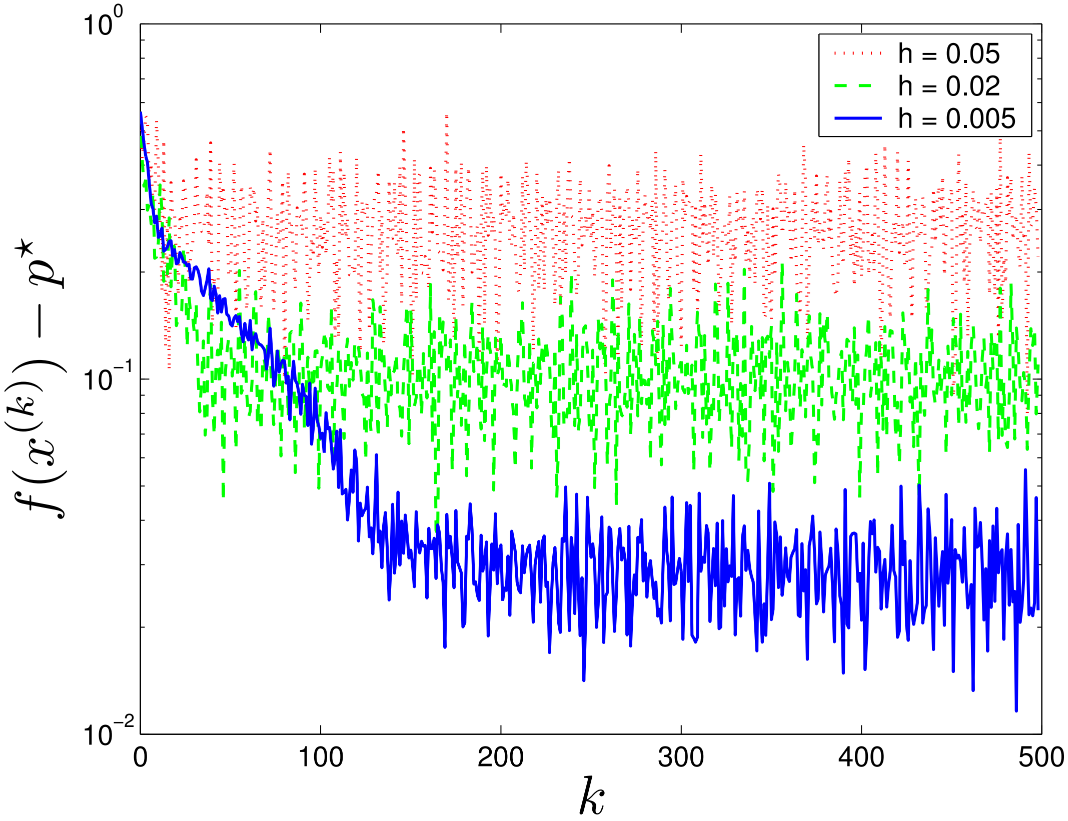
constant step length, $h = 0.05,\ 0.02, 0.005$



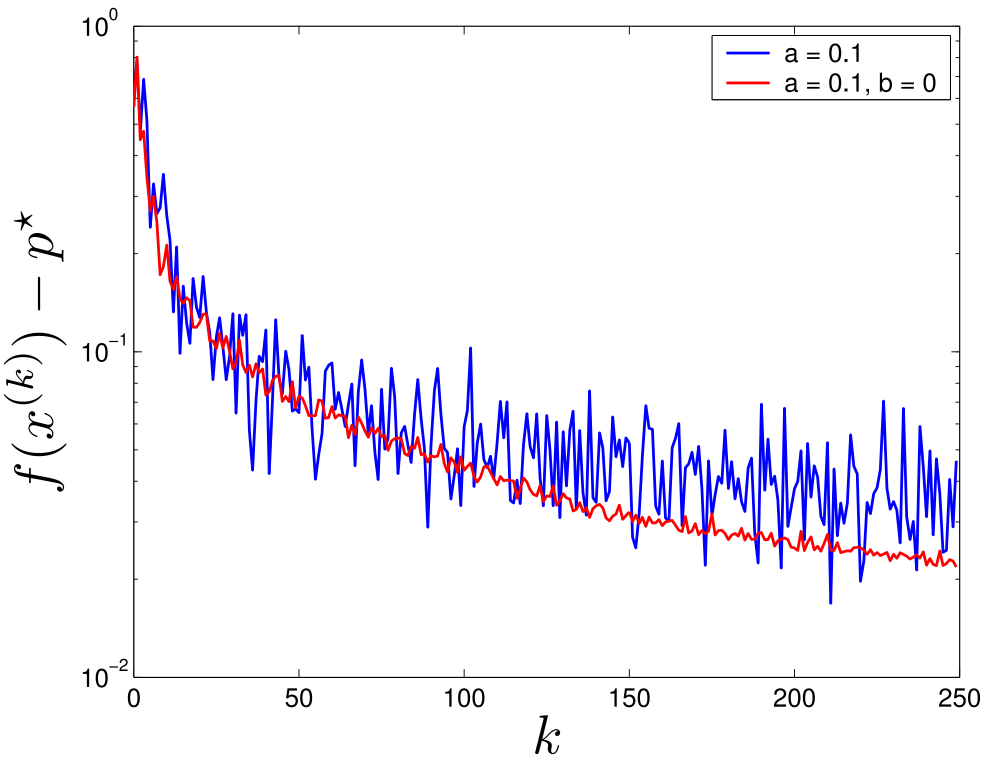
$f_{\mathrm{best}}^{(k)} - p^\star$ and upper bound, constant step length $h = 0.02$



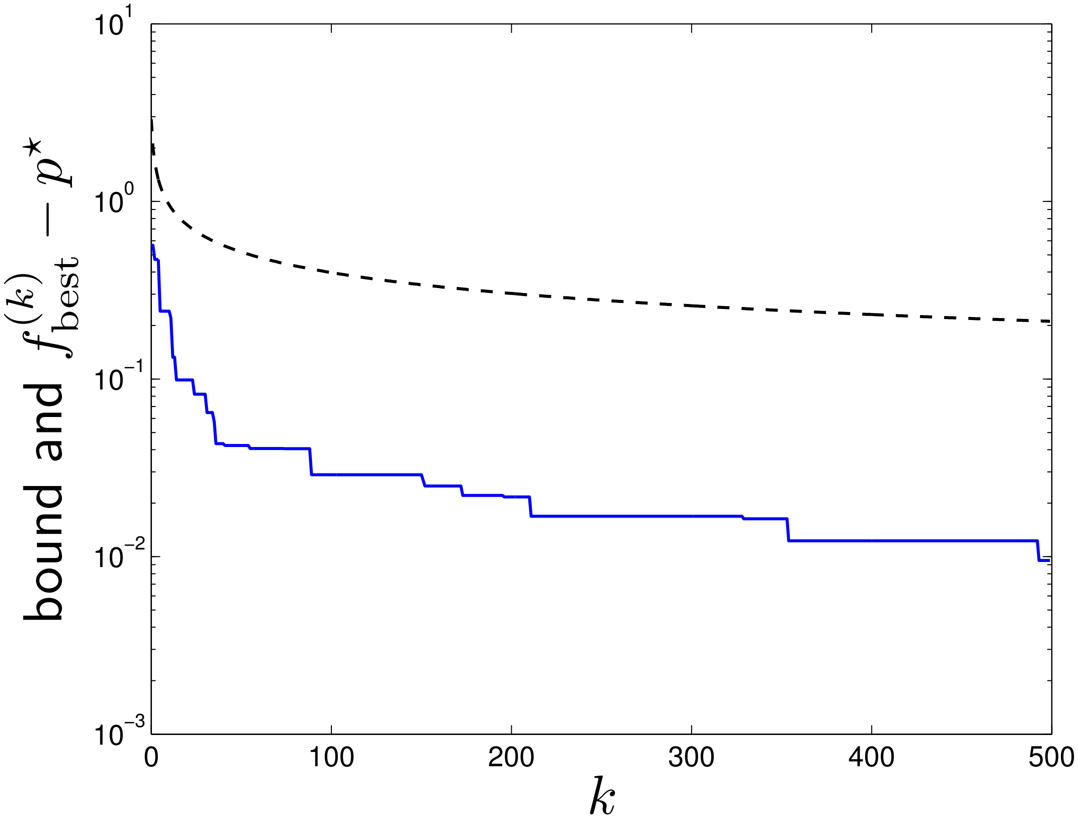
constant step size $h = 0.05,\ 0.02,\ 0.005$



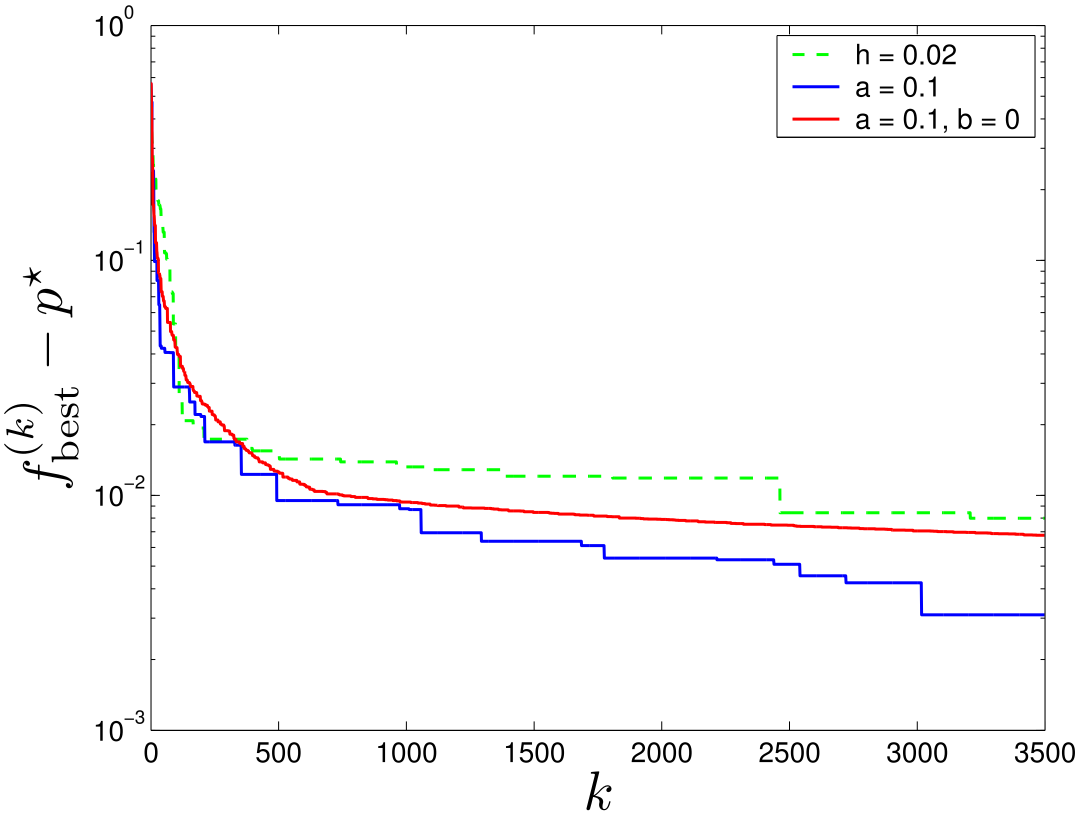
diminishing step rule $\alpha = 0.1/\sqrt{k}$ and square summable step size rule $\alpha = 0.1/k$.



$f_{\rm best}^{(k)} - p^\star$ and upper bound, diminishing step size rule $\alpha = 0.1/\sqrt{k}$



constant step length $h = 0.02$, diminishing step size rule $\alpha = 0.1/\sqrt{k}$, and square summable step rule $\alpha = 0.1/k$



Projected subgradient method

solves constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C}, \end{array}$$

where $f : \mathbf{R}^n \to \mathbf{R}$, $\mathcal{C} \subseteq \mathbf{R}^n$ are convex

projected subgradient method is given by

$$x^{(k+1)} = P(x^{(k)} - \alpha_k g^{(k)}),$$

P is (Euclidean) projection on $\mathcal{C}$, and $g^{(k)} \in \partial f(x^{(k)})$

same convergence results:

- for constep step size, converges to neighborhood of optimal (for f differentiable and h small enough, converges)
- for diminishing nonsummable step sizes, converges

key idea: projection does not increase distance to $x^\star$

approximate projected subgradient: P only needs to satisfy

$$P(u) \in \mathcal{C}, \qquad \|P(u) - z\|_2 \leq \|u - z\|_2 \text{ for any } z \in \mathcal{C}$$

Projected subgradient for dual problem

(convex) primal:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{array}$$

solve dual problem

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

via projected subgradient method:

$$\lambda^{(k+1)} = \left(\lambda^{(k)} - \alpha_k h\right)_+, \qquad h \in \partial(-g)(\lambda^{(k)})$$

Subgradient of negative dual function

assume f_0 is strictly convex, and denote, for $\lambda \succeq 0$,

$$x^*(\lambda) = \underset{z}{\operatorname{argmin}} \left(f_0(z) + \lambda_1 f_1(z) + \cdots + \lambda_m f_m(z) \right)$$

so $g(\lambda) = f_0(x^*(\lambda)) + \lambda_1 f_1(x^*(\lambda)) + \cdots + \lambda_m f_m(x^*(\lambda))$

a subgradient of $-g$ at λ is given by $h_i = -f_i(x^*(\lambda))$

projected subgradient method for dual:

$$x^{(k)} = x^*(\lambda^{(k)}), \qquad \lambda_i^{(k+1)} = \left(\lambda_i^{(k)} + \alpha_k f_i(x^{(k)}) \right)_+$$

note:

- primal iterates $x^{(k)}$ are not feasible, but become feasible in limit
- subgradient of $-g$ directly gives the violation of the primal constraints
- dual function values $g(\lambda^{(k)})$ converge to $p^\star$

Example: Optimal network flow

- connected directed graph with n links, p nodes
- variable x_j denotes the flow or traffic on arc j (can be < 0)
- given external source (or sink) flow s_i at node i, $\mathbf{1}^T s = 0$
- flow conservation: $Ax = s$, where $A \in \mathbf{R}^{p \times n}$ is *node incidence matrix*
- $\phi_j : \mathbf{R} \to \mathbf{R}$ convex flow cost function for link j

optimal (single commodity) network flow problem:

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{n} \phi_j(x_j) \\ \text{subject to} & Ax = s \end{array}$$

Dual network flow problem

Langrangian is

$$\begin{aligned} L(x,\nu) &= \sum_{j=1}^{n} \phi_j(x_j) + \nu^T(s - Ax) \\ &= \sum_{j=1}^{n} (\phi_j(x_j) - \Delta\nu_j x_j) + \nu^T s \end{aligned}$$

- we interpret ν_i as *potential* at node i
- $\Delta\nu_j$ denotes *potential difference* across link j

dual function is

$$
\begin{aligned}
q(\nu) = \inf_x L(x, \nu) &= \sum_{j=1}^{n} \inf_{x_j} \left(\phi_j(x_j) - \Delta\nu_j x_j\right) + \nu^T s \\
&= -\sum_{j=1}^{n} \phi_j^*(\Delta\nu_j) + \nu^T s
\end{aligned}
$$

(ϕ_j^* is conjugate function of ϕ_j)

dual network flow problem:

$$
\text{maximize} \quad q(\nu) = -\textstyle\sum_{j=1}^{n} \phi_j^*(\Delta\nu_j) + \nu^T s
$$

Optimal network flow via dual

assume ϕ_i strictly convex, and denote

$$x_j^*(\Delta\nu_j) = \underset{x_j}{\operatorname{argmin}} \left(\phi_j(x_j) - \Delta\nu_j x_j\right)$$

if $\nu^\star$ is optimal solution of the dual network flow problem,

$$x_j^\star = x_j^*(\Delta\nu_j^\star)$$

is optimal flow

Electrical network analogy

- electrical network with node incidence matrix A, nonlinear resistors in branches
- variable x_j is the current flow in branch j
- source s_i is external current injected at node i (must sum to zero)
- flow conservation equation $Ax = s$ is Kirkhoff Current Law (KCL)
- dual variables are node potentials; $\Delta\nu_j$ is jth branch voltage
- branch current-voltage characteristic is $x_j = x_j^*(\Delta\nu_j)$

then, current and potentials in circuit are optimal flows and dual variables

Subgradient of negative dual function

a subgradient of the negative dual function $-q$ at ν is

$$g = Ax^*(\Delta\nu) - s$$

ith component is $g_i = a_i^T x^*(\Delta\nu) - s_i$, which is *flow excess* at node i

Subgradient method for dual

subgradient method applied to dual can be expressed as:

$$\begin{aligned} x_j &:= x_j^*(\Delta\nu_j) \\ g_i &:= a_i^T x - s_i \\ \nu_i &:= \nu_i - \alpha g_i \end{aligned}$$

interpretation:

- optimize each flow, given potential difference, without regard for flow conservation
- evaluate flow excess
- update potentials to correct flow excesses

Example: Minimum queueing delay

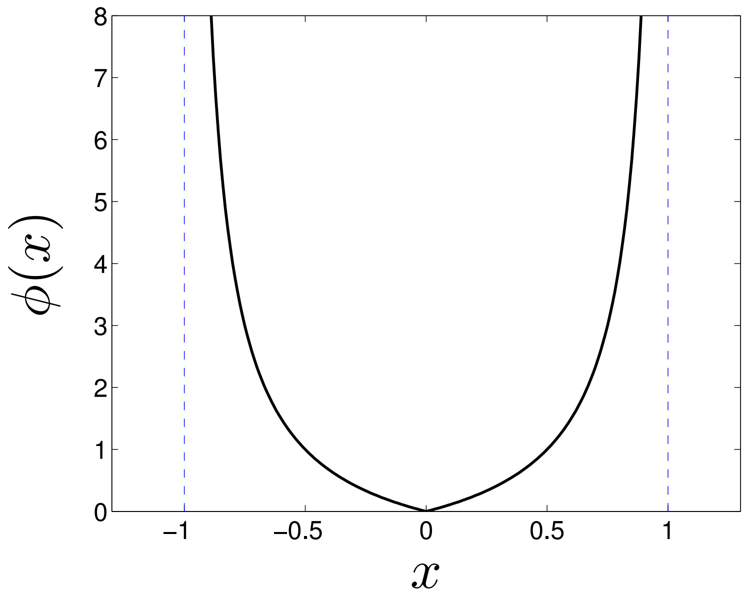
flow cost function

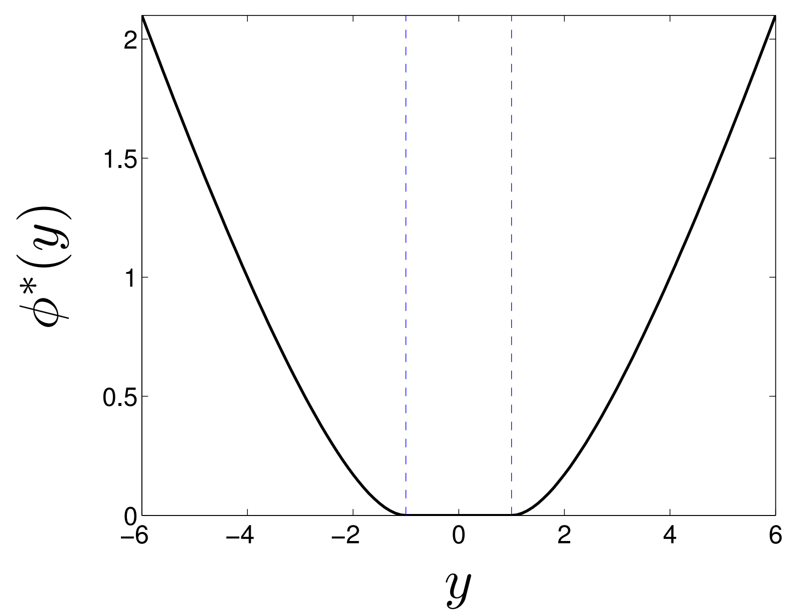
$$\phi_j(x_j) = \frac{|x_j|}{c_j - |x_j|}, \qquad \mathbf{dom}\, \phi_j = (-c_j, c_j)$$

where $c_j > 0$ are given *link capacities*
($\phi_j(x_j)$ gives expected waiting time in queue with exponential arrivals at rate x_j, exponential service at rate c_j)
conjugate is

$$\phi_j^*(y) = \begin{cases} 0 & |y| \leq 1/c_j \\ \left(\sqrt{|c_j y|} - 1\right)^2, & |y| > 1/c_j \end{cases}$$

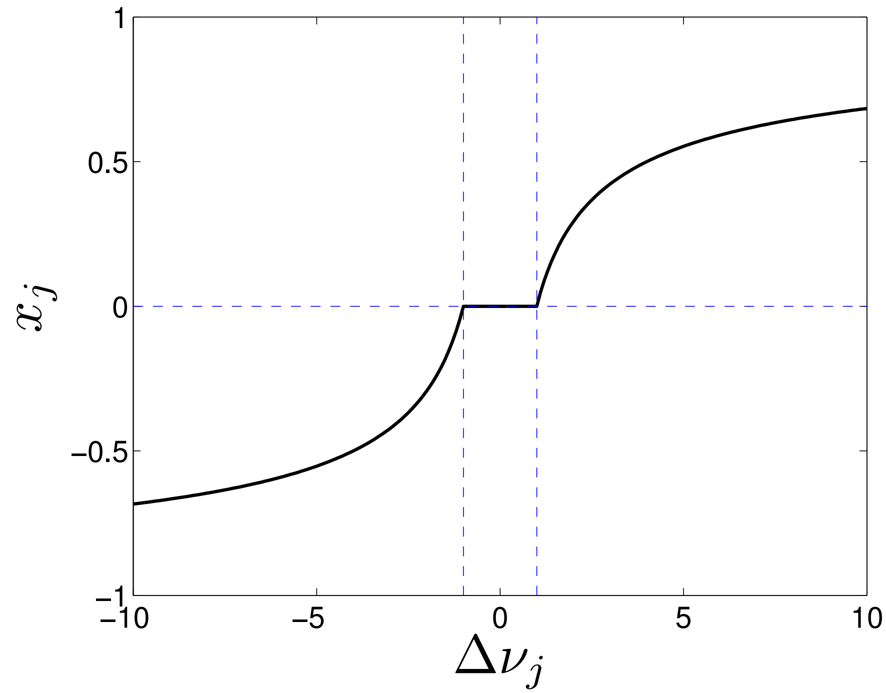
cost function $\phi(x)$ (left) and its conjugate $\phi^*(y)$ (right), $c = 1$





(note conjugate is differentiable)

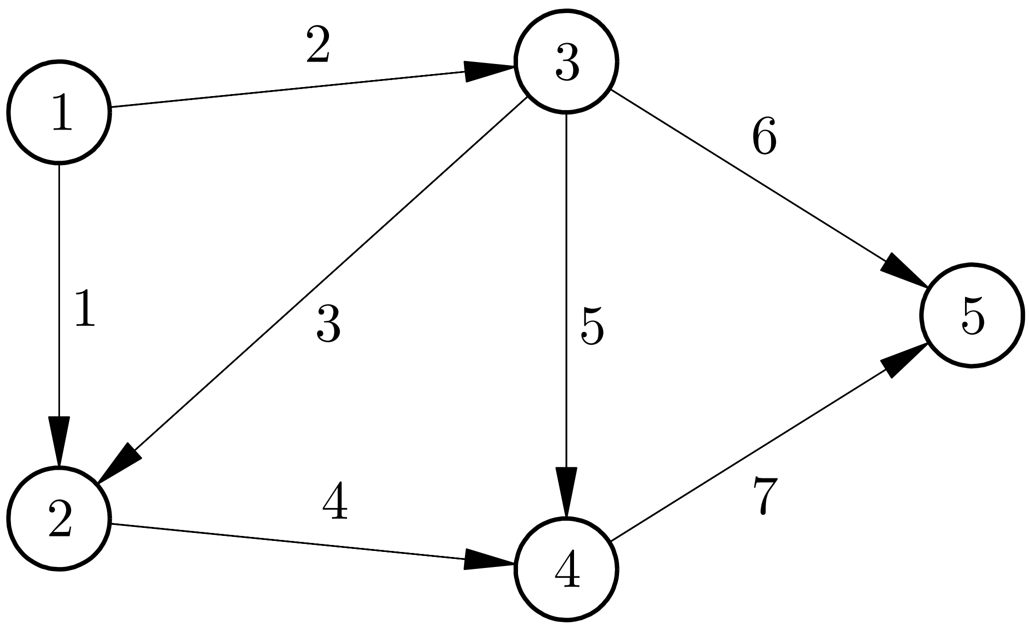
$x_j^*(\Delta\nu_j)$, for $c_j = 1$



gives flow as function of potential difference across link

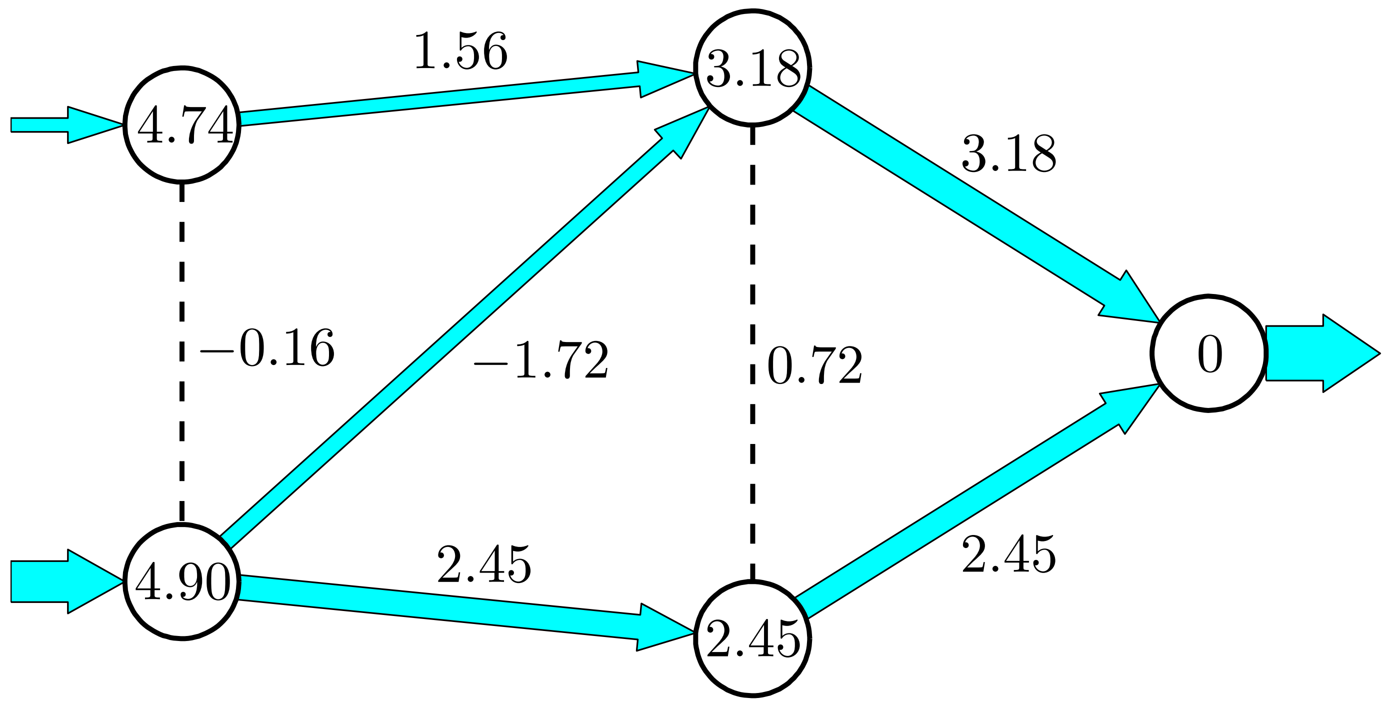
A specific example

network with 5 nodes, 7 links, capacities $c_j = 1$



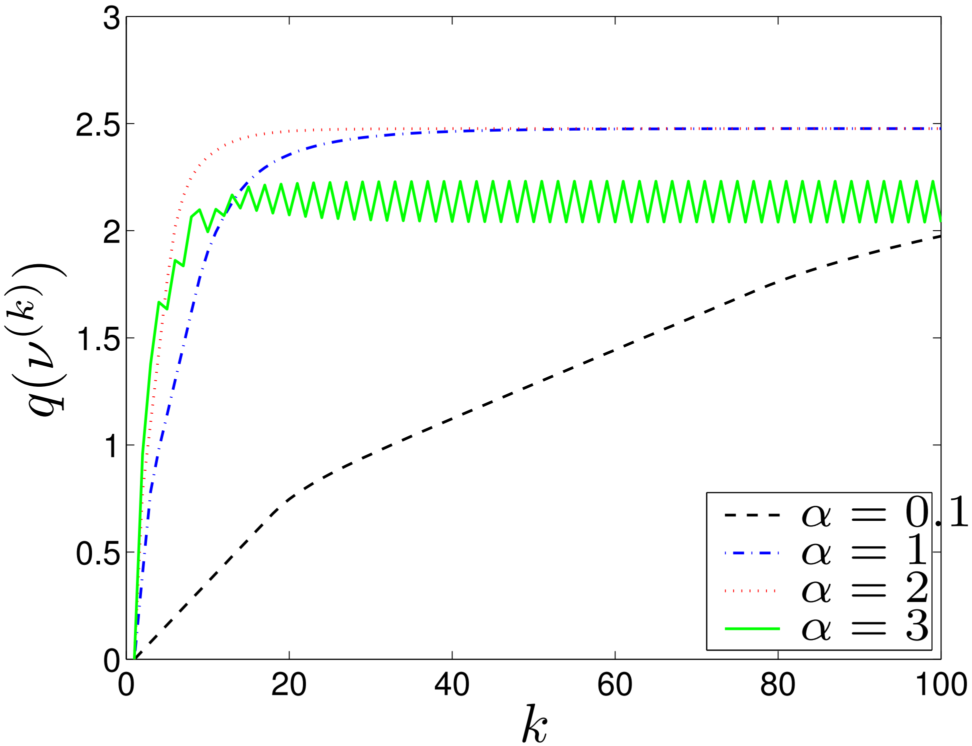
Optimal flow

optimal flows shown as width of arrows; optimal dual variables shown in nodes; potential differences shown on links



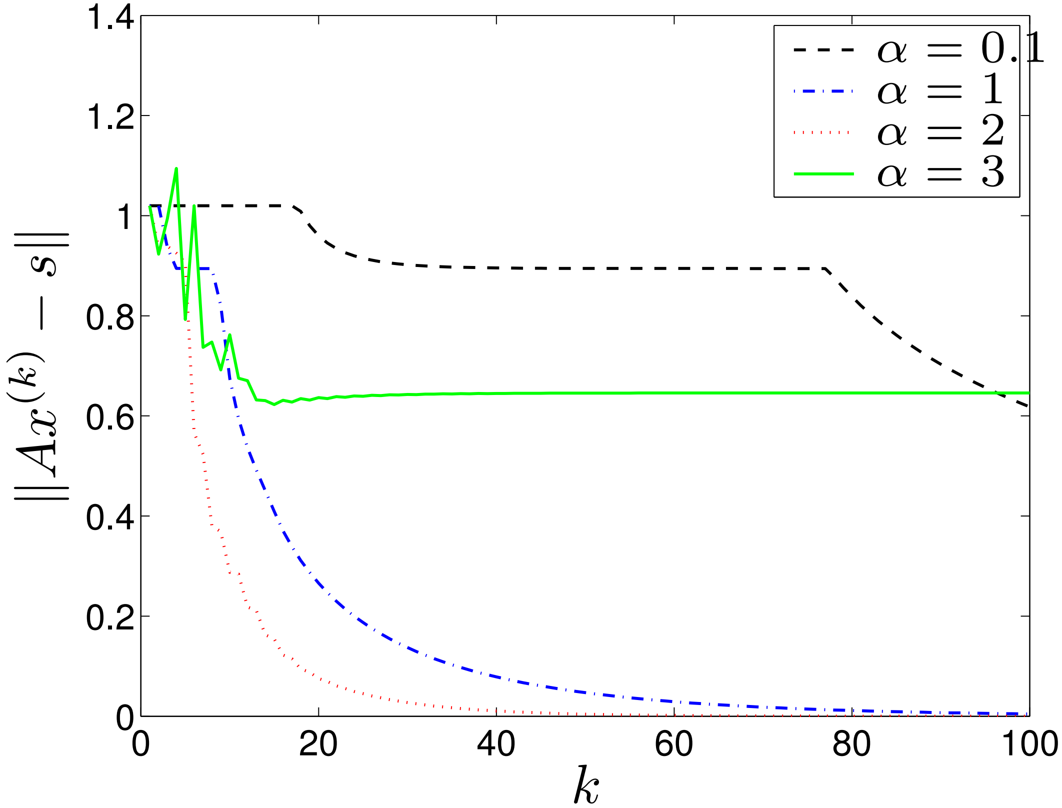
Convergence of dual function

constant stepsize rule, $\alpha = 0.1,\ 1,\ 2,\ 3$

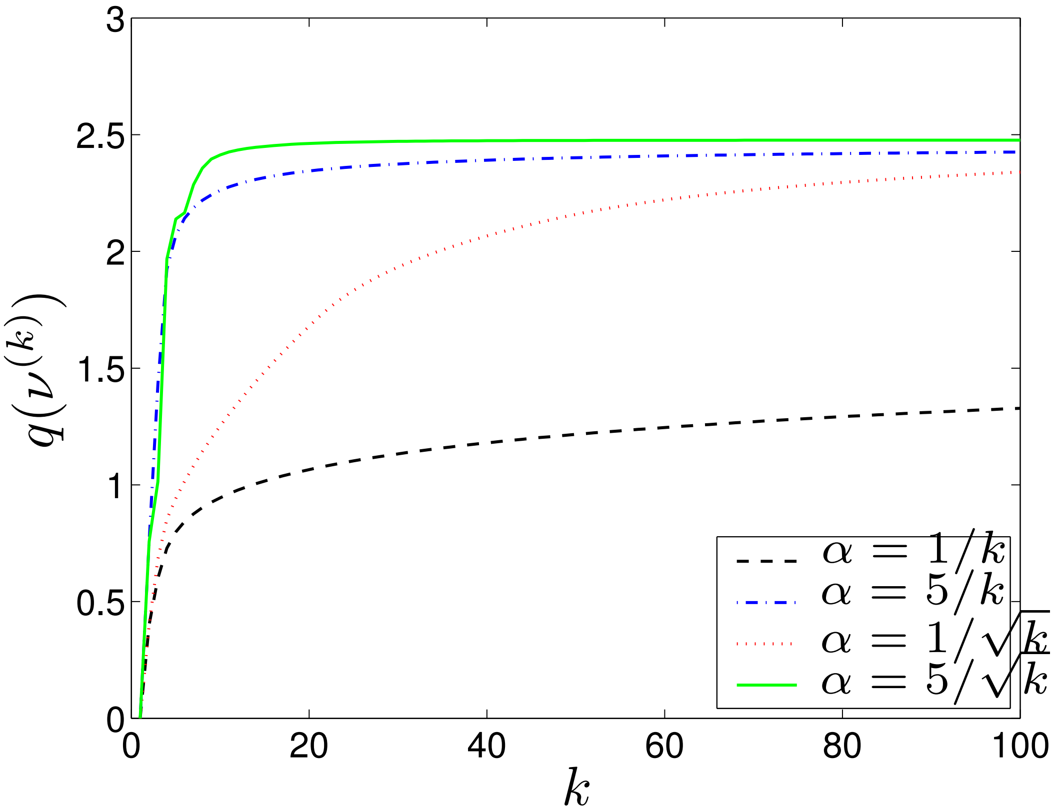


for $\alpha = 1,\ 2$, converges to $p^\star = 2.48$ in around about 40 iterations

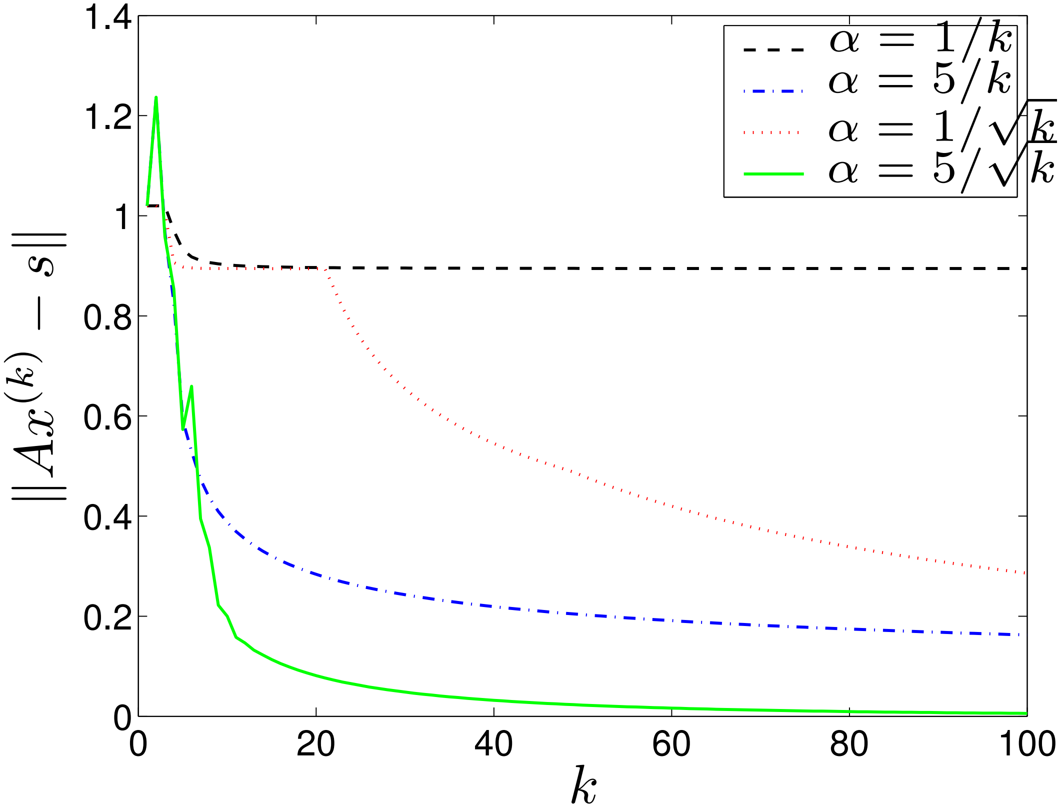
Convergence of primal residual



convergence of dual function, nonsummable diminishing stepsize rules

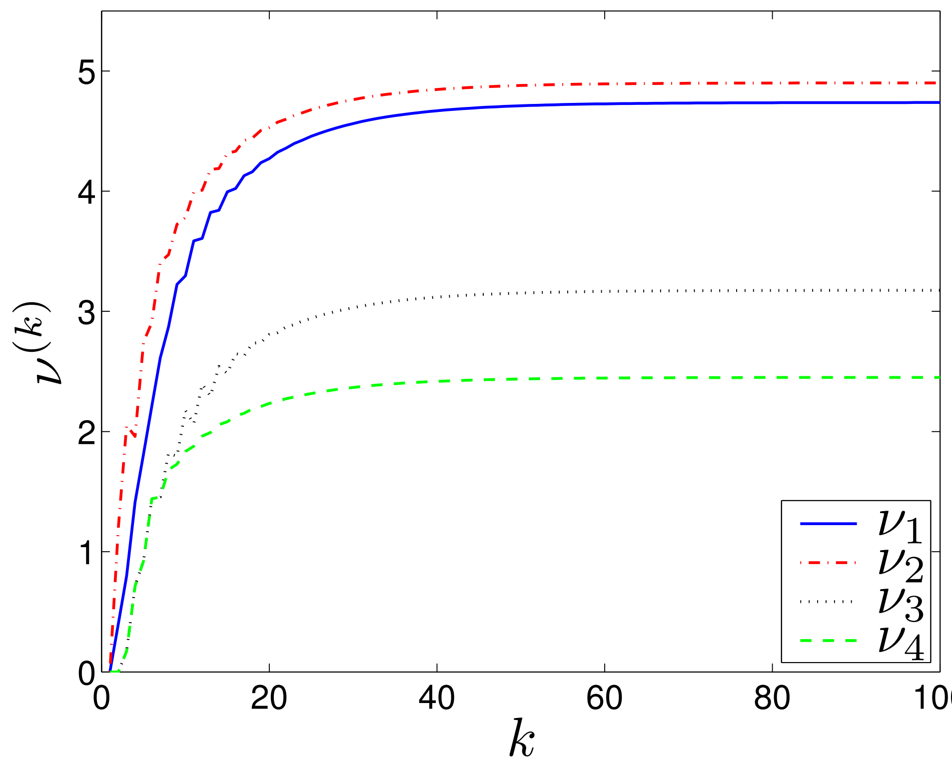


convergence of primal residual, nonsummable diminishing stepsize rules



Convergence of dual variables

$\nu^{(k)}$ versus iteration number k, constant stepsize rule $\alpha = 2$



(ν_5 is fixed as zero)