9. Dual decomposition and dual algorithms

- dual gradient ascent
- example: network rate control
- dual decomposition and the proximal gradient method
- examples with simple dual prox-operators
- alternating minimization method

Dual methods

convex problem with linear constraints and its dual

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array} \qquad\qquad \begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

dual function can be expressed in terms of conjugate of f:

$$\begin{aligned} g(\lambda, \nu) &= \inf_x \left(f(x) + (G^T\lambda + A^T\nu)^T x - h^T\lambda - b^T\nu \right) \\ &= -h^T\lambda - b^T\nu - f^*(-G^T\lambda - A^T\nu) \end{aligned}$$

potential advantages of solving the dual when using 1-st order methods

- dual is unconstrained or has simple constraints
- dual decomposes into smaller problems

(Sub-)gradients of conjugate function

assume $f : \mathbf{R}^n \to \mathbf{R}$ is closed, convex with conjugate

$$f^*(y) = \sup_x \left(y^T x - f(x)\right)$$

- $x \in \partial f^*(y)$ if and only if x maximizes $y^T x - f(x)$ (p. 6-10)

- if f is strictly convex, then f^* is differentiable on $\textbf{int dom}\, f^*$ and

$$\nabla f^*(y) = \operatorname*{argmax}_x \left(y^T x - f(x)\right)$$

- if f is strongly convex with parameter $\mu > 0$, then f^* is differentiable, $\textbf{dom}\, f^* = \mathbf{R}^n$, and

$$\|\nabla f^*(y) - \nabla f^*(x)\|_2 \leq \frac{1}{\mu}\|x - y\|_2$$

(see p. 8-7)

Dual gradient method

primal problem: (for simplicity, only equality constraints)

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

dual problem: maximize $g(\nu)$ where

$$g(\nu) = \inf_x \left(f(x) + (Ax - b)^T \nu \right)$$

dual ascent: solve dual by (sub-)gradient method (t is stepsize)

$$x^+ = \operatorname*{argmin}_x \left(f(x) + \nu^T Ax \right), \qquad \nu^+ = \nu + t(Ax^+ - b)$$

- sometimes referred to as Uzawa's method
- of interest if calculation of x^+ is inexpensive

Dual decomposition

convex problem with separable objective

$$\begin{array}{ll} \text{minimize} & f_1(x_1) + f_2(x_2) \\ \text{subject to} & G_1 x_1 + G_2 x_2 \preceq h \end{array}$$

constraint is *complicating or coupling* constraint

dual problem (master problem)

$$\begin{array}{ll} \text{maximize} & g_1(\lambda) + g_2(\lambda) - h^T \lambda \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

where $g_j(\lambda) = \inf \left(f_j(x) + \lambda^T G_j x \right) = -f_j^*(-G_j^T \lambda)$

can be solved by (sub-)gradient projection (if $\lambda \succeq 0$ is the only constraint)

subproblem: to calculate $g_j(\lambda)$ and a (sub-)gradient, solve problem

$$\text{minimize (over } x_j) \quad f_j(x_j) + \lambda^T G_j x_j$$

- optimal value is $g_j(\lambda)$
- if $\hat{x}_j$ solves the subproblem, then $-G_j\hat{x}_j$ is a subgradient of $-g_j$ at λ

dual subgradient projection method

- solve two unconstrained (and independent) subproblems

$$x_j^+ = \operatorname*{argmin}_{x_j} \left(f_j(x_j) + \lambda^T G_j x_j\right), \quad j = 1, 2$$

- make projected subgradient update of λ

$$\lambda^+ = \left(\lambda + t(G_1 x_1^+ + G_2 x_2^+ - h)\right)_+$$

($u_+ = \max\{u, 0\}$, componentwise)

interpretation: price coordination

- $p = 2$ units in the system; unit j selects variable x_j
- constraints are limits on shared resources; λ_i is price of resource i
- dual update $\lambda_i^+ = (\lambda_i - ts_i)_+$ depends on slacks $s = h - G_1 x_1 - G_2 x_2$
  - increases price λ_i if resource is over-used ($s_i < 0$)
  - decreases price λ_i if resource is under-used ($s_i > 0$)
  - never lets price get negative

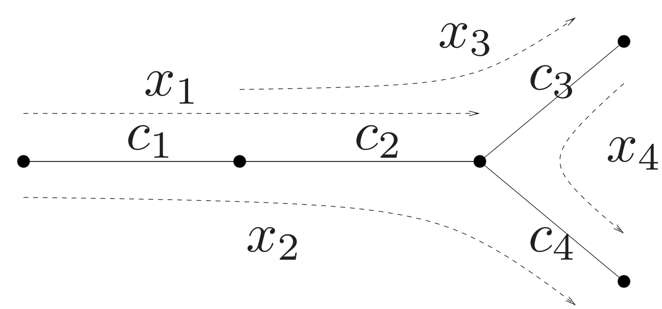
distributed architecture

- central node 0 sets price λ
- peripheral node j sets x_j

Example: network rate control

- n flows (with fixed routes) in a network with m links
- variable $x_j \geq 0$ denotes rate of flow j
- utility function for flow j is $U_j : \mathbf{R} \to \mathbf{R}$, concave, increasing

capacity constraints



- traffic y_i on link i is sum of flows passing through it
- $y = Rx$, where R is the routing matrix

$$R_{ij} = \begin{cases} 1 & \text{flow } j \text{ passes through link } i \\ 0 & \text{otherwise} \end{cases}$$

- link capacity constraint: $y \preceq c$

$$
\begin{array}{ll}
\text{maximize} & U(x) = \sum_{j=1}^n U_j(x_j) \\
\text{subject to} & Rx \preceq c
\end{array}
$$

a convex problem; dual decomposition gives decentralized method

Lagrangian (for minimizing $-U$)

$$L(x, \lambda) = -U(x) + \lambda^T(Rx - c) = -\lambda^T c + \textstyle\sum_{j=1}^n (-U_j(x_j) + x_j r_j^T \lambda)$$

- λ_i is the price (per unit flow) for using link i
- $r_j^T \lambda$ is the sum of prices along route j (r_j is jth column of R)

dual function

$$g(\lambda) = -\lambda^T c + \sum_{j=1}^n \inf_{x_j} (-U_j(x_j) + x_j r_j^T \lambda) = -\lambda^T c - \sum_{j=1}^n (-U_j)^*(-r_j^T \lambda)$$

(Sub-)gradients of dual function

$$g(\lambda) = -\lambda^T c - \sum_{j=1}^{n} \sup_{x_j} \left(U_j(x_j) - x_j r_j^T \lambda\right)$$

- subgradient of $-g(\lambda)$

$$c - R\bar{x} \in \partial(-g)(\lambda) \quad \text{where} \quad \bar{x}_j = \operatorname{argmax}\left(U_j(x_j) - x_j r_j^T \lambda\right)$$

  if U_j is strictly concave, this is a gradient

- $r_j^T \lambda$ is the sum of link prices along route j

- $c - R\bar{x}$ is vector of link capacity margins for flow $\bar{x}$

Dual decomposition rate control algorithm

given initial link price vector $\lambda \succ 0$ (e.g., $\lambda = \mathbf{1}$)

repeat

1. sum link prices along each route: calculate $\Lambda_j = r_j^T \lambda$
2. optimize flows (separately) using flow prices:

$$x_j^+ := \operatorname{argmax}\left(U_j(x_j) - \Lambda_j x_j\right)$$

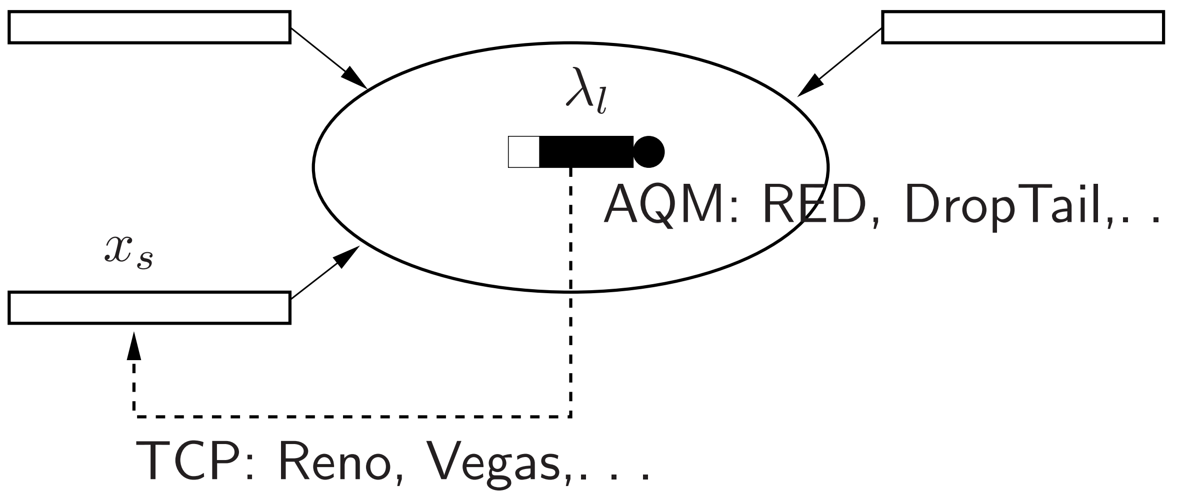
3. calculate link capacity margins $s := c - Rx$
4. update link prices: (t is the step size)

$$\lambda := (\lambda - ts)_+$$

decentralized: links only need to know the flows that pass through them; flows only need to know prices on links they pass through

TCP/AQM congestion control

a large class of internet congestion control mechanisms can be interpreted as distributed algorithms that solve NUM and its dual



x_s: source rate, updated by TCP (Transmission Control Protocol)

λ_l: link congestion measure, or 'price', updated by AQM (Active Queue Management)
e.g., TCP Reno uses packet loss as congestion measure, TCP Vegas uses queueing delay

refs: [Kelly,et al,'98];[Low,Lapsley'99];. . .

Outline

- dual gradient ascent
- example: network rate control
- **dual decomposition and dual proximal gradient method**
- examples with simple dual prox-operators
- alternating minimization method

First-order dual methods

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array} \qquad \begin{array}{ll} \text{maximize} & -f^*(-G^T\lambda - A^T\nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

can apply different algorithms to the dual:

subgradient method: slow convergence

gradient method: requires differentiable f

- in many applications f^* is not differentiable, has a nontrivial domain
- f^* can be smoothed by adding a small strongly convex term to f

proximal gradient method: dual costs split in two terms

- first term is differentiable; second term has an inexpensive prox-operator

Composite structure in the dual

primal problem with separable objective

$$\begin{array}{ll} \text{minimize} & f(x) + h(y) \\ \text{subject to} & Ax + By = b \end{array}$$

(later we consider general problem with inequality constraints)

dual problem

$$\text{maximize} \quad -f^*(-A^T\nu) - h^*(-B^T\nu) - b^T\nu$$

has the composite structure required for the proximal gradient method if

- f is strongly convex, hence ∇f^* is Lipschitz continuous
- prox-operator of $h^*(-B^T\nu)$ is cheap (closed form or efficient algorithm)

Example: regularized norm approximation

$$\text{minimize} \quad f(x) + \|Ax - b\|$$

f is strongly convex with parameter μ; $\|\cdot\|$ is any norm

(reformulated) problem and dual

$$\begin{array}{ll} \text{minimize} & f(x) + \|y\| \\ \text{subject to} & y = Ax - b \end{array} \qquad\qquad \begin{array}{ll} \text{maximize} & b^T z - f^*(A^T z) \\ \text{subject to} & \|z\|_* \leq 1 \end{array}$$

- gradient of dual cost is Lipschitz continuous with parameter $\|A\|_2^2/\mu$
- for most norms, projection on norm ball is inexpensive

dual gradient projection step (with $C = \{v \mid \|v\|_* \leq 1\}$)

$$z^+ = P_C\left(z + t(b - A\nabla f^*(A^T z))\right)$$

where $\nabla f^*(A^T z) = \mathrm{argmin}_x\,(f(x) - z^T Ax)$

gradient projection algorithm: choose initial z and repeat

$$\begin{array}{rcl} \hat{x} & := & \underset{x}{\mathrm{argmin}}\,(f(x) - z^T Ax) \\ z & := & P_C(z + t(b - A\hat{x})) \end{array}$$

- step size t: constant or from backtracking line search
- can also use accelerated gradient projection algorithm

Example: regularized nuclear norm approximation

$$\text{minimize} \quad \frac{1}{2}\|x - a\|_2^2 + \|A(x) - B\|_*$$

$\|\cdot\|_*$ is nuclear norm and $A : \mathbf{R}^n \to \mathbf{R}^{p\times q}$ with $A(x) = \sum_{i=1}^n x_i A_i$

gradient projection: choose initial Z and repeat

$$\begin{aligned} \hat{x}_i &:= a_i + \mathbf{tr}(A_i^T Z), \quad i = 1, \ldots, n \\ Z &:= P_C(Z + t(B - A(\hat{x}))) \end{aligned}$$

- $\hat{x}$ is minimizer of $(1/2)\|x - a\|_2^2 - \sum_i x_i \,\mathbf{tr}(A_i^T Z)$
- C is unit ball for matrix norm $\|V\| = \sigma_{\max}(V)$
- to find $P_C(V)$, replace σ_i by $\min\{\sigma_i, 1\}$ in SVD of V

Example: dual decomposition

$$\text{minimize} \quad f(x) + \sum_{i=1}^{p} \|B_i x\|_2$$

with f strongly convex, $B_i \in \mathbf{R}^{m_i \times n}$

reformulated problem

$$\begin{array}{ll} \text{minimize} & f(x) + \sum_{i=1}^{p} \|y_i\|_2 \\ \text{subject to} & y_i = B_i x, \quad i = 1, \ldots, p \end{array}$$

objective is separable, but not strictly convex

dual problem

$$\begin{array}{ll} \text{maximize} & -f^*(\sum_{i=1}^{p} B_i^T z_i) \\ \text{subject to} & \|z_i\|_2 \leq 1, \quad i = 1, \ldots, p \end{array}$$

dual gradient projection step (with $C_i = \{v \in \mathbf{R}_i^m \mid \|v\|_2 \leq 1\}$)

$$z_i^+ = P_{C_i}\left(z_i - tB_i \nabla f^*(\sum_{i=1}^p B_i^T z_i)\right), \quad i = 1, \ldots, p$$

algorithm: choose initial z_i and repeat

$$\begin{aligned} z &:= \sum_{i=1}^p B_i^T z_i \\ \hat{x} &:= \operatorname*{argmin}_x \left(f(x) - z^T x\right) \quad (= \nabla f^*(z)) \\ z_i &:= P_{C_i}(z_i - tB_i\hat{x}), \quad i = 1, \ldots, p \end{aligned}$$

- updates of z_i are independent
- if f is separable, primal update decomposes into independent subproblems

Minimization over intersection of convex sets

$$\begin{array}{ll}\text{minimize} & f(x) \\ \text{subject to} & x \in C_1 \cap \ldots \cap C_m\end{array}$$

- f strongly convex; C_i closed, convex with inexpensive projector
- example: $f(x) = \|x - a\|_2^2$ gives projection of a on intersection

reformulation: introduce auxiliary variables x_i

$$\begin{array}{ll}\text{minimize} & f(x) + I_{C_1}(x_1) + \ldots + I_{C_m}(x_m) \\ \text{subject to} & x_1 = x, \; \ldots, x_m = x\end{array}$$

dual problem

$$\text{maximize} \quad -f^*(z_1 + \ldots + z_m) - h_1(z_1) - \ldots - h_m(z_m)$$

$h_i(z) = \sup_{x \in C_i}(-z^T x)$ is support function of C_i at $-z$

dual proximal gradient step

$$z_i^+ = \mathbf{prox}_{th_i}(z_i - t\nabla f^*(z_1 + \ldots + z_m)), \quad i = 1, \ldots, m$$

prox-operator of h_i can be expressed in terms of projection on C_i

$$\mathbf{prox}_{th_i}(u) = u + tP_{C_i}(-u/t)$$

dual proximal gradient algorithm: choose initial $z_1, \ldots, z_m$ and repeat

$$\begin{aligned} \hat{x} &:= \operatorname*{argmin}_x \left(f(x) - (z_1 + \ldots + z_m)^T x\right) \\ z_i &:= z_i + t\left(P_{C_i}(\hat{x} - \frac{1}{t}z_i) - \hat{x}\right), \quad i = 1, \ldots, m \end{aligned}$$

can take $t = \mu/m$ (μ is strong convexity parameter of f)

Outline

- dual gradient ascent
- network rate control (utility maximization)
- dual decomposition and dual proximal gradient method
- examples with simple dual prox-operators
- **alternating minimization method**

Prox-operator of partial dual

$$\begin{array}{ll} \text{minimize} & f(x) + h(y) \\ \text{subject to} & Ax + By = b \end{array} \qquad\qquad \begin{array}{ll} \text{minimize} & -f^*(-A^T\nu) - F(\nu) \end{array}$$

- F is negative of a 'partial dual function'

$$\begin{aligned} F(\nu) &= b^T\nu + h^*(-B^T\nu) \\ &= -\inf_x \left(h(y) + \nu^T(By - b)\right) \end{aligned}$$

- prox-operator of F is defined as

$$\mathbf{prox}_{tF}(\nu) = \operatorname*{argmin}_v \left(F(v) + \frac{1}{2t}\|v - \nu\|_2^2 \right)$$

Primal expression for prox-operator

- by definition, $v = \mathbf{prox}_{tF}(\nu)$ is the minimizer v of

$$b^T v + h^*(-B^T v) + \frac{1}{2t}\|v - \nu\|_2^2$$

- this is the dual of the problem (with variables y, z)

$$\text{maximize} \quad -h(y) - \nu^T z - \tfrac{t}{2}\|z\|_2^2, \quad \text{subject to} \quad By - b = z$$

- primal and dual optimal solutions are related by $v = \nu + t(By - b)$

conclusion: primal method for computing $v = \mathbf{prox}_{tF}(\nu)$

$$\hat{y} = \operatorname{argmin}\left(h(y) + \nu^T(By - b) + \frac{t}{2}\|By - b\|_2^2\right), \quad v = \nu + t(B\hat{y} - b)$$

$\hat{y}$ minimizes **augmented Lagrangian** (Lagrangian + quadratic penalty)

Alternating minimization method

$$\begin{array}{ll} \text{minimize} & f(x) + h(y) \\ \text{subject to} & Ax + By = b \end{array} \qquad\qquad \begin{array}{ll} \text{minimize} & -f^*(-A^T\nu) - F(\nu) \end{array}$$

f strongly convex; h convex, not necessarily strictly

dual proximal gradient step

$$\nu^+ = \mathbf{prox}_{tF}(\nu + tA\nabla f^*(-A^T\nu))$$

- $\hat{x} = \nabla f^*(-A^T\nu)$ is minimizer of $f(x) + \nu^T Ax$
- $\mathbf{prox}_{tF}(\nu + tA\hat{x}) = \nu + t(A\hat{x} + B\hat{y} - b)$ where $\hat{y}$ minimizes

$$h(y) + (\nu + tA\hat{x})^T(By - b) + \frac{t}{2}\|By - b\|_2^2$$

algorithm: choose initial ν and repeat

$$
\begin{aligned}
\hat{x} &:= \operatorname*{argmin}_x \left(f(x) + \nu^T A x\right) \\
\hat{y} &:= \operatorname*{argmin}_y \left(h(y) + \nu^T B y + \frac{t}{2}\|A\hat{x} + By - b\|_2^2\right) \\
\nu &:= \nu + t(A\hat{x} + B\hat{y} - b)
\end{aligned}
$$

- alternating minimization of
  - Lagrangian (step 1)
  - augmented Lagrangian (step 2)
- step 3 is proximal gradient update for the dual problem
- as a variation, can use accelerated proximal gradient method

General problem with separable objective

$$\begin{array}{ll} \text{minimize} & f(x) + h(y) \\ \text{subject to} & Ax + By = b \\ & Cx + Dy \preceq d \end{array}$$

f strongly convex

dual problem

$$\text{maximize} \quad -f^*(-C^T\lambda - A^T\nu) - F(\lambda, \nu)$$

where

$$F(\lambda, \nu) = \begin{cases} d^T\lambda + b^T\nu + h^*(-D^T\lambda - B^T\nu), & \lambda \succeq 0 \\ +\infty, & \text{otherwise} \end{cases}$$

we derive expressions for the prox-operator of F

Proximal operator of partial dual function

definition: $(u, v) = \mathbf{prox}_{tF}(\lambda, \nu)$ is the solution of

$$\text{minimize} \quad F(u, v) + \frac{1}{2t}(\|u - \lambda\|_2^2 + \|v - \nu\|_2^2)$$

equivalent expression

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \lambda \\ \nu \end{bmatrix} + t \begin{bmatrix} D\hat{y} + \hat{s} - d \\ B\hat{y} - b \end{bmatrix}$$

where $\hat{y}$, $\hat{s}$ solve

$$\begin{array}{ll} \text{minimize} & h(y) + \lambda^T(Dy + s) + \nu^T By + \frac{1}{2t}(\|Dy + s - d\|_2^2 + \|By - b\|_2^2) \\ \text{subject to} & s \succeq 0 \end{array}$$

proof: follows from the duality between the problems

$$\begin{array}{ll} \text{minimize}_{x,s,w,z} & h(y) + \lambda^T w + \nu^T z + \frac{1}{2t}(\|w\|_2^2 + \|z\|_2^2) \\ \text{subject to} & Dy + s - d = w \\ & By - b = z \\ & s \succeq 0 \end{array}$$

and

$$\begin{array}{ll} \text{maximize}_{u,v} & -d^T u - b^T v - h^*(-D^T u - B^T v) - \frac{1}{2t}(\|u - \lambda\|_2^2 + \|v - \nu\|_2^2) \\ \text{subject to} & u \succeq 0 \end{array}$$

- at the optimum,

$$\lambda + t(Dy + s - d) = u, \quad \nu + t(By - b) = v$$

- by definition the optimal (u, v) is the proximal operator $\mathbf{prox}_{tF}(\lambda, \nu)$

Alternating minimization method

choose initial λ, ν and repeat

1. compute the minimizer $\hat{x}$ of the Lagrangian

$$f(x) + (A^T\nu + C^T\lambda)^T x$$

2. compute the minimizers $\hat{y}$, $\hat{s}$ of the augmented Lagrangian

$$h(y) + \lambda^T(Dy + s) + \nu^T By + \frac{t}{2}(\|C\hat{x} + Dy + s - d\|_2^2 + \|A\hat{x} + By - b\|_2^2)$$

subject to $s \succeq 0$

3. dual update

$$\lambda := \lambda + t(C\hat{x} + D\hat{y} - \hat{s} - d), \quad \nu := \nu + t(A\hat{x} + B\hat{y} - b)$$

as a variation, can use a fast proximal gradient update

References and sources

- L. Vandenberghe, *Lecture notes for EE236C - Optimization Methods for Large-Scale Systems* (Spring 2011), UCLA.
- S. Boyd, course notes for EE364b, Convex Optimization II (the rate control example)
- D.P. Bertsekas and J.N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods (1989)
- F. Kelly, A. Maulloo, D. Tan, Rate control in communication networks: shadow prices, proportional fairness and stability, J. Operation Research Society, 49 (1998).
- A. Beck and M. Teboulle, Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems, IEEE Transactions on Image Processing (2009)
- P. Tseng, Applications of a splitting algorithm to decomposition in convex programming and variational inequalities, SIAM J. Control and Optimization (1991)
- P. Tseng, Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming, Mathematical Programming (1990) Dual proximal gradient method 10-2