Dual Ascent

Lecturer: Aarti Singh
Co-instructor: Pradeep Ravikumar

Convex Optimization 10-725/36-725

Summary of Duality

Lagrangian duality to derive lower bound on primal objective:

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x), \quad u \geq 0$$

$$f(x) \geq L(x, u, v) \quad \forall x \text{ feasible}, u \geq 0, v$$

$$f^* \geq \min_{x \in C} L(x, u, v) \geq \min_{x} L(x, u, v) =: g(u, v) \quad \forall u \geq 0, v$$

Note: Procedure applies to non-convex problems as well

Summary of Duality

Lagrangian duality to derive lower bound on primal objective:

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x), \quad u \geq 0$$

$$f(x) \geq L(x, u, v) \quad \forall x \text{ feasible}, u \geq 0, v$$

$$f^* \geq \min_{x \in C} L(x, u, v) \geq \min_{x} L(x, u, v) =: g(u, v) \quad \forall u \geq 0, v$$

Note: Procedure applies to non-convex problems as well

Primal problem		Dual problem	
$\min_x$	$f(x)$	$\max_{u,v}$	$g(u, v)$
subject to	$h_i(x) \leq 0, i = 1, \ldots, m$	subject to	$u \geq 0$
	$\ell_j(x) = 0, j = 1, \ldots, r$		

Since

$$L(x, u, v) = f(x) + u^\top h(x) + v^\top \ell(x)$$

we have

$$\max_{u \geq 0, v} L(x, u, v) = \begin{cases} f(x) & h(x) \leq 0,\ \ell(x) = 0 \text{ (i.e. } x \text{ feasible)} \\ \infty & \text{otherwise} \end{cases}$$

Hence, we get:

Primal problem	Dual problem
$\min_x \max_{u \geq 0, v} L(x, u, v)$	$\max_{u \geq 0, v} \min_x L(x, u, v)$

Weak duality

$$f^* \geq g^*$$

Note: Holds even for non-convex problems

Strong duality

$$f^* = g^*$$

Note: Holds for convex problems under Slater's condition: There exists at least one strictly feasible $x \in \mathbb{R}^n$, meaning

$$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots \ell_r(x) = 0$$

An important refinement: strict inequalities only need to hold over functions h_i that are not affine

Uses of duality

Pros:

- Optimal dual objective gives a lower bound (or sometimes same value as) on the optimal primal objective
- Dual problem has as many variables as constraints in primal problem - maybe easier to solve
- Dual problem often has simpler constraints - maybe easier to solve
- Dual problem is convex (concave maximization) even if primal is not - maybe easier to solve

Uses of duality

Pros:

- Optimal dual objective gives a lower bound (or sometimes same value as) on the optimal primal objective
- Dual problem has as many variables as constraints in primal problem - maybe easier to solve
- Dual problem often has simpler constraints - maybe easier to solve
- Dual problem is convex (concave maximization) even if primal is not - maybe easier to solve
- Duality gap can be used as a stopping criterion (next)

Uses of duality

Pros:

- Optimal dual objective gives a lower bound (or sometimes same value as) on the optimal primal objective
- Dual problem has as many variables as constraints in primal problem - maybe easier to solve
- Dual problem often has simpler constraints - maybe easier to solve
- Dual problem is convex (concave maximization) even if primal is not - maybe easier to solve
- Duality gap can be used as a stopping criterion (next)
- KKT conditions can be used to understand (and under strong duality, derive) primal solution; algorithms based on KKT conditions (next)

Uses of duality

Pros:

- Optimal dual objective gives a lower bound (or sometimes same value as) on the optimal primal objective
- Dual problem has as many variables as constraints in primal problem - maybe easier to solve
- Dual problem often has simpler constraints - maybe easier to solve
- Dual problem is convex (concave maximization) even if primal is not - maybe easier to solve
- Duality gap can be used as a stopping criterion (next)
- KKT conditions can be used to understand (and under strong duality, derive) primal solution; algorithms based on KKT conditions (next)
- Algorithms based on dual problem, e.g. dual ascent (next)

Uses of duality

Cons:

- May be difficult to evaluate the dual (requires unconstrained minimization of Lagrangian)
- Dual function is often non-differentiable
- Dual optimal solution (u^*, v^*) in general does not yield primal optimal solution x^* (unless strong duality holds)

Duality gap

Given primal feasible x and dual feasible u, v, the quantity

$$f(x) - g(u, v)$$

is called the duality gap between x and u, v. Note that

$$f(x) - f^\star \leq f(x) - g(u, v)$$

so if the duality gap is zero, then x is primal optimal (and similarly, u, v are dual optimal)

From an algorithmic viewpoint, provides a stopping criterion: if $f(x) - g(u, v) \leq \epsilon$, then we are guaranteed that $f(x) - f^\star \leq \epsilon$

Very useful, especially in conjunction with iterative methods ...

KKT conditions

Consider a general primal optimization problem (no assumptions of convexity or differentiability).

The KKT(Karush-Kuhn-Tucker) conditions are

- $0 \in \partial\left(f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x) \right)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \leq 0,\ \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \geq 0$ for all i (dual feasibility)

Sufficiency

General (no assumptions of convexity or differentiability): If

- $x^* = \arg\min_x L(x, u^*, v^*) \Leftrightarrow 0 \in \partial L(x^*, u^*, v^*)$ (stationarity)
- x^* is primal feasible
- $u^* \geq 0$ i.e. dual feasible
- $u_i^* = 0 \; \forall i \notin A(x^*) := \{i : h_i(x^*) = 0\}$
  $\Leftrightarrow u_i^* \cdot h_i(x^*) = 0 \; \forall i$ (complementary slackness)

then x^* is global minimum of the problem.

Sufficiency

General (no assumptions of convexity or differentiability): If

- $x^* = \arg\min_x L(x, u^*, v^*) \Leftrightarrow 0 \in \partial L(x^*, u^*, v^*)$ (stationarity)
- x^* is primal feasible
- $u^* \geq 0$ i.e. dual feasible
- $u_i^* = 0 \; \forall i \notin A(x^*) := \{i : h_i(x^*) = 0\}$
  $\Leftrightarrow u_i^* \cdot h_i(x^*) = 0 \; \forall i$ (complementary slackness)

then x^* is global minimum of the problem.

Note: (u^*, v^*) are also dual optimal.

$$\begin{aligned} g(u^\star, v^\star) &= \min_x \{f(x) + u^{\star\top} h(x) + v^{\star\top} \ell(x)\} \\ &= f(x^\star) + u^{\star\top} h(x^\star) + v^{\star\top} \ell(x^\star) = f(x^\star) \end{aligned}$$

where the first equality holds from stationarity, and the second holds from complementary slackness and primal feasibility

KKT conditions are sufficient for primal and dual optimality.

Alternate sufficiency conditions:

If problem is convex and differentiable, stationarity condition becomes

$$0 = \nabla_x L(x^*, u^*, v^*)$$

and corresponding KKT conditions are sometimes called first-order sufficiency conditions.

[1] for both equality and inequality constraints, see DB book Prop 3.3.2

Alternate sufficiency conditions:

If problem is convex and differentiable, stationarity condition becomes

$$0 = \nabla_x L(x^*, u^*, v^*)$$

and corresponding KKT conditions are sometimes called first-order sufficiency conditions.

If problem is twice differentiable but not necessarily convex (discussed earlier for equality constraints only[1]), then if x^*, v^* satisfy

$$0 = \nabla_x L(x^*, v^*),$$

$$0 = \nabla_v L(x^*, v^*) \quad \Leftrightarrow \quad \ell(x^*) = 0,$$

$$y^\top \nabla^2_{xx} L(x^*, u^*, v^*) y > 0 \quad \forall y \neq 0, \nabla \ell(x^*)^\top y = 0$$

then it is guaranteed that x^* is a local minimum. These are called second-order sufficiency conditions.

[1] for both equality and inequality constraints, see DB book Prop 3.3.2

Necessity

KKT conditions are necessary for primal and dual optimality under strong duality.

Let $x^\star$ and $u^\star, v^\star$ be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater's condition). Then

- x^* is primal feasible
- u^*, v^* are dual feasible

Necessity

KKT conditions are necessary for primal and dual optimality under strong duality.

Let $x^\star$ and $u^\star, v^\star$ be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater's condition). Then

- x^* is primal feasible
- u^*, v^* are dual feasible

Also, zero duality gap implies

$$\begin{aligned} f(x^\star) &= g(u^\star, v^\star) \\ &= \min_x \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x) \\ &\leq f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \leq f(x^\star) \end{aligned}$$

where last inequality holds since x^* is primal feasible.

In other words, all previous inequalities are actually equalities. This implies:

- the point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$.

$$0 \in \partial L(x^*, u^*, v^*) \quad \text{(stationarity)}$$

- $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i

$$u_i^* \cdot h_i(x^*) = 0 \ \forall i \quad \text{(complementary slackness)}$$

In other words, all previous inequalities are actually equalities. This implies:

- the point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$.

$$0 \in \partial L(x^*, u^*, v^*) \quad \text{(stationarity)}$$

- $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i

$$u_i^* \cdot h_i(x^*) = 0 \ \forall i \quad \text{(complementary slackness)}$$

i.e. If $x^\star$ and $u^\star, v^\star$ are primal and dual solutions with zero duality gap, then they satisfy KKT conditions.

Note: The sufficiency and necessity statements don't assume anything about convexity or differentiability

Note: The necessity condition just presented require strong duality to hold, but do not require regularity assumptions

Alternate sufficiency conditions under regularity assumptions:

Let x^* be a local minimum and a regular point. Then there exist unique Lagrange multiplier vectors u^*, v^* such that

$$0 = \nabla_x L(x^*, u^*, v^*),$$

$$u_i^* \geq 0, i = 1, \ldots, m; \quad u_i^* = 0 \ \forall i \notin A(x^*)$$

$$y^\top \nabla_{xx}^2 L(x^*, u^*, v^*) y \geq 0 \quad \forall y \in V(x^*)$$

where

$$V(x^*) = \{y : \nabla h_i(x^*)^\top y = 0 \text{ for } i \in A(x^*), \nabla \ell(x^*)^\top y = 0\}$$

Characterizing primal using dual

Recall that under strong duality, the KKT conditions are necessary for optimality. Thus, if the dual is solved exactly to yield u^*, v^*, then the primal solution must minimize $L(x, u^*, v^*)$.

- Generally, this reveals a characterization of primal solutions
- In particular, if this is satisfied uniquely (i.e., above problem has a unique minimizer), then the corresponding point must be the primal optimal solution.

but can also yield other solutions that are primal infeasible.

Characterizing primal using dual

Recall that under strong duality, the KKT conditions are necessary for optimality. Thus, if the dual is solved exactly to yield u^*, v^*, then the primal solution must minimize $L(x, u^*, v^*)$.

- Generally, this reveals a characterization of primal solutions
- In particular, if this is satisfied uniquely (i.e., above problem has a unique minimizer), then the corresponding point must be the primal optimal solution.

but can also yield other solutions that are primal infeasible.

Example: One way to establish sparsity of lasso solution and conditions under which it holds is via constructing a set of primal and dual candidate solutions (certificate) that satisfy KKT conditions, and observing the conditions which allow the primal to be sparse [Wainwright'09].

Algorithms based on KKT conditions

Since the KKT conditions are sufficient for primal (and dual) optimality, we can try to solve for primal x and dual variables u, v that satisfy KKT conditions. These will then be primal and dual optimal due to sufficiency.

The KKT conditions can be thought of as a system of nonlinear equations that can be solved approximately via Newton's method. We saw two methods inspired by this idea:

- Barrier method
- Primal-dual method

Both solve for perturbed KKT conditions (where complementary slackness is perturbed) that are easier to solve than standard KKT conditions.

Algorithms based on dual problem

Since dual problem is always convex (concave maximization) irrespective of primal, we can use the methods for convex minimization we have learnt so far.

Algorithms based on dual problem

Since dual problem is always convex (concave maximization) irrespective of primal, we can use the methods for convex minimization we have learnt so far.

Key challenge: Differentiability of Lagrange dual function $g(u, v)$

- Whenever $L(x, u, v)$ is minimized over a unique $x_{u,v}$ for any given (u, v), then g is differentiable.
- This holds, for example, if f is strictly convex and h is affine.
- But in general, this often does not hold. In particular, whenever there is duality gap, the dual function is not differentiable at every dual optimal solution.

Algorithms based on dual problem

Since dual problem is always convex (concave maximization) irrespective of primal, we can use the methods for convex minimization we have learnt so far.

Key challenge: Differentiability of Lagrange dual function $g(u, v)$

- Whenever $L(x, u, v)$ is minimized over a unique $x_{u,v}$ for any given (u, v), then g is differentiable.
- This holds, for example, if f is strictly convex and h is affine.
- But in general, this often does not hold. In particular, whenever there is duality gap, the dual function is not differentiable at every dual optimal solution.

Algorithms for dual problems:

- Differentiable - Dual gradient ascent (next)
- Non-differentiable - Dual subgradient ascent (next), Cutting plane, Decomposition methods

Dual ascent

Since dual problem is always convex (concave maximization) irrespective of primal, we can use gradient or sub-gradient ascent on the dual variables.

Let x' be a minimizer of $L(x, u', v')$ for given $u' \geq 0, v'$. Then

$\begin{bmatrix} h(x') \\ \ell(x') \end{bmatrix}$ is a (sub)gradient of g at $\begin{bmatrix} u' \\ v' \end{bmatrix}$ because $\forall u, v$

Dual ascent

Since dual problem is always convex (concave maximization) irrespective of primal, we can use gradient or sub-gradient ascent on the dual variables.

Let x' be a minimizer of $L(x, u', v')$ for given $u' \geq 0, v'$. Then

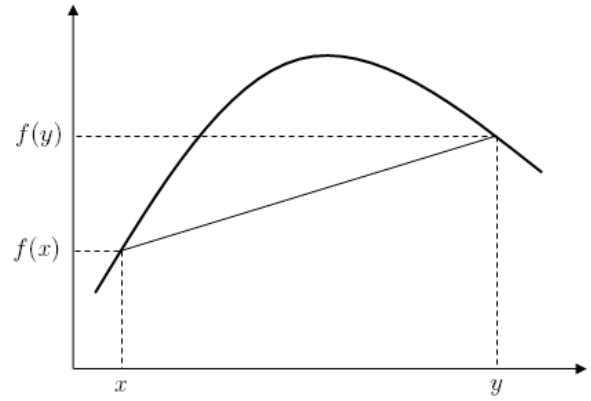
$\begin{bmatrix} h(x') \\ \ell(x') \end{bmatrix}$ is a (sub)gradient of g at $\begin{bmatrix} u' \\ v' \end{bmatrix}$ because $\forall u, v$

$$\begin{aligned}
g(u,v) &= \min_x L(x,u,v) \\
&= \min_x f(x) + u^\top h(x) + v^\top \ell(x) \\
&\leq f(x') + u^\top h(x') + v^\top \ell(x') \\
&= f(x') + u'^\top h(x') + (u-u')^\top h(x') \\
&\quad + v'^\top \ell(x') + (v-v')^\top \ell(x') \\
&= g(u',v') + (u-u')^\top h(x') + (v-v')^\top \ell(x')
\end{aligned}$$

Last step follows since x' is a minimizer of $L(x, u', v')$.

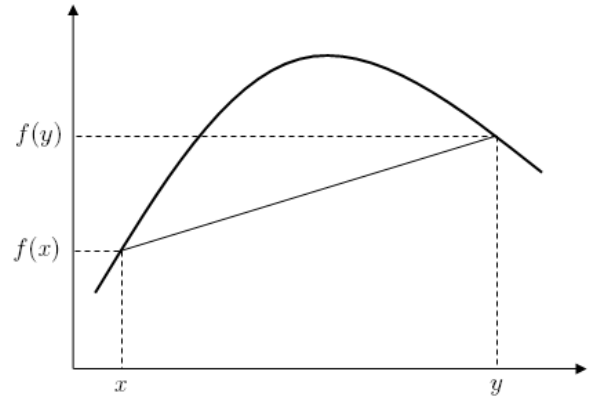
Recall: Subgradient of a *concave* function f at x is any s s.t.

$$f(y) \leq f(x) + s^\top (y - x) \quad \forall y$$



Recall: Subgradient of a *concave* function f at x is any s s.t.

$$f(y) \leq f(x) + s^\top (y - x) \quad \forall y$$



Dual (sub)gradient ascent method

- Start
  with an initial dual guess $u^{(0)} \geq 0, v^{(0)}$.
- Repeat for $k = 1, 2, 3, \ldots$

$$x^{(k)} \in \operatorname*{argmin}_x \; f(x) + (u^{(k-1)})^\top h(x) + (v^{(k-1)})^\top \ell(x)$$
$$u^{(k)} = \max\{u^{(k-1)} + t_k h(x^{(k)}), 0\}$$
$$v^{(k)} = v^{(k-1)} + t_k \ell(x^{(k)})$$

Step sizes $t_k, k = 1, 2, 3, \ldots$ are chosen in standard ways
Proximal gradients and acceleration can be applied as they would usually

Method of multipliers as dual ascent

Recall Method of Multipliers: Solve sequence of unconstrained minimization of Augmented Lagrangian

$$x^{(k)} = \arg\min_x \; L_{c^{(k)}}(x, \lambda^{(k)})$$

where for equality constrained problem ($\min_x f(x)$ s.t. $h(x) = 0$)

$$L_{c^{(k)}}(x, \lambda^{(k)}) = f(x) + {\lambda^{(k)}}^\top h(x) + \frac{c^{(k)}}{2}\|h(x)\|^2$$

and using the following multiplier update:

$$\lambda^{(k+1)} = \lambda^{(k)} + c^{(k)} h(x^{(k)}).$$

This is precisely dual ascent for the augmented problem!

Gradient vs Subgradient descent/ascent

- Subgradient may not be a direction of ascent at (u, v) where dual function g is non-differentiable, so we take best iterate so far:

$$g((u^{(k)}, v^{(k)})_{\text{best}}) = \max_{i=0,\ldots k} g(u^{(i)}, v^{(i)})$$

Gradient vs Subgradient descent/ascent

- Subgradient may not be a direction of ascent at (u, v) where dual function g is non-differentiable, so we take best iterate so far:

$$g((u^{(k)}, v^{(k)})_{\text{best}}) = \max_{i=0,\ldots k} g(u^{(i)}, v^{(i)})$$

- The subgradient makes an angle < 90 with all ascent directions at (u, v)

$$f(y) \leq f(x) + s^\top (y - x) \quad \forall y \quad \Rightarrow \quad 0 < s^\top (y - x) \quad \forall f(y) > f(x)$$

Gradient vs Subgradient descent/ascent

- Subgradient may not be a direction of ascent at (u, v) where dual function g is non-differentiable, so we take best iterate so far:

$$g((u^{(k)}, v^{(k)})_{\text{best}}) = \max_{i=0,\ldots k} g(u^{(i)}, v^{(i)})$$

- The subgradient makes an angle < 90 with all ascent directions at (u, v)

$$f(y) \leq f(x) + s^\top (y - x) \quad \forall y \quad \Rightarrow \quad 0 < s^\top (y - x) \quad \forall f(y) > f(x)$$

This implies that a small move from (u, v) in the direction of any subgradient at u, v decreases the distance to any maximizer of g. To see this, let $v_{k+1} = v_k + t_k s_k$. Then

$$\|v_{k+1} - v^*\|^2 \quad = \quad \|v_k - v^*\|^2 + t_k^2 \|s_k\|^2 + 2 t_k s_k^\top (v_k - v^*)$$

Since $g(v_k) \leq g(v^*)$, we have

$$\|v_{k+1} - v^*\| \leq \|v_k - v^*\| \quad \forall 0 < t_k < 2(g(v^*) - g(v_k))/\|s_k\|^2$$

Step size choices

- Fixed step sizes: $t_k = t$ all $k = 1, 2, 3, \ldots$
- Diminishing step sizes: choose to meet conditions

$$\sum_{k=1}^{\infty} t_k^2 < \infty, \quad \sum_{k=1}^{\infty} t_k = \infty,$$

  i.e., square summable but not summable

  Important that step sizes go to zero, but not too fast

Other options too, but important difference to gradient descent: step sizes are typically pre-specified, not adaptively computed

Dual decomposition

Consider

$$\min_x \sum_{i=1}^B f_i(x_i) \quad \text{subject to} \quad \sum_{i=1}^B h_{ij}(x_i) \leq 0 \quad j = 1, \ldots, m$$

Here $x = (x_1, \ldots x_B) \in \mathbb{R}^n$ divides into B blocks of variables, with each $x_i \in \mathbb{R}^{n_i}$.

Simple but powerful observation, in calculation of (sub)gradient, is that the minimization decomposes into B separate problems:

$$x^+ \in \operatorname*{argmin}_x \sum_{i=1}^B (f_i(x_i) + u^\top h_i(x_i))$$

$$\iff x_i^+ \in \operatorname*{argmin}_{x_i} \; f_i(x_i) + u^T h_i(x_i), \quad i = 1, \ldots B$$

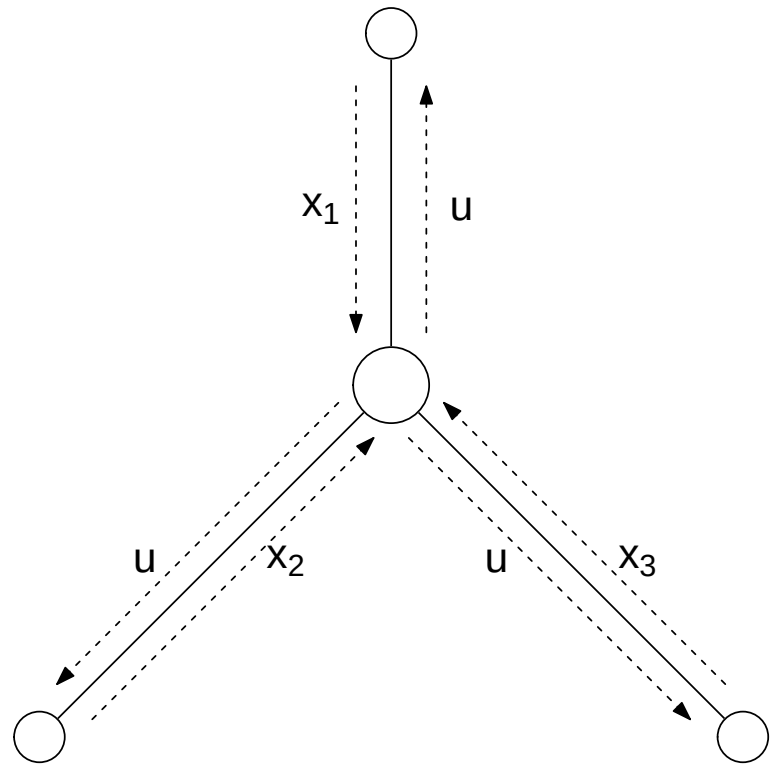
Dual decomposition algorithm: repeat for $k = 1, 2, 3, \ldots$

$$x_i^{(k)} \in \operatorname*{argmin}_{x_i} \; f_i(x_i) + (u^{(k-1)})^T h_i(x_i), \quad i = 1, \ldots B$$

$$u^{(k)} = \max \left\{ u^{(k-1)} + t_k \left(\sum_{i=1}^{B} h_i(x_i^{(k)}) \right), 0 \right\}$$

Can think of these steps as:

- Broadcast: send u to each of the B processors, each optimizes in parallel to find x_i
- Gather: collect $h_i(x_i)$ from each processor, update the global dual variable u



Price coordination interpretation (Vandenberghe):

- Have B units in a system, each unit chooses its own decision variable x_i (how to allocate its goods)
- There are m resources. Constraints are limits on shared resources ($\sum_{i=1}^{B} h_{ij}(x)$ is constraint on resource j), each component of dual variable u_j is price of resource j
- Dual update:

$$u_j^+ = (u_j + t\xi_j)_+, \quad j = 1, \ldots m$$

  where $\xi_j = \sum_{i=1}^{B} h_{ij}(x_i)$ are slacks

  - Increase price u_j if resource j is over-utilized, $\xi_j > 0$
  - Decrease price u_j if resource j is under-utilized, $\xi_j < 0$
  - Never let prices get negative