Decomposition techniques for distributed optimization: 
a tutorial overview

Mikael Johansson
KTH – Stockholm – Sweden

( with a voice gone bust ☹ )

Aim of these lectures

"To present some of the key techniques for decomposition and 
distributed optimization in a coherent and comprehensible manner"

Focus on understanding, not all the details
  – each lecture could be a full-semester course
  – you will have to work with the material yourself!

Focus on fundamentals
  – many techniques date back to 60’s-80’s, ...
  – but some are very recent, and research frontier is not far away

References at end of presentation (will be posted on-line later this week)
Why distributed optimization

Optimization on a “Google scale”
– information processing on huge data sets

Coordination and control of large-scale systems
– power and water distribution
– vehicle coordination and planning
– sensor, social, and data networks

Theoretical foundation for communication protocol design
– Internet congestion control
– scheduling and power control in wireless systems

Example: water distribution

Coordinated control of water distribution in city of Barcelona (WIDE)
Example: multi-agent coordination

Cooperate to find jointly optimal controls and rendez-vous point

\[
\text{minimize} \quad \sum_{i \in V} f_i(\theta) \\
\text{subject to} \quad \theta \in \Theta
\]

where

\[
f_i(\theta) = \min_{\theta} \sum_{t=0}^{T} (x_t - \theta)^T Q (x_t - \theta) + u_t^T R u_t \\
\text{s.t.} \quad x_{t+1} = A x_t + B u_t, \quad t = 0, \ldots, T - 1
\]

Example: communication protocol design

Understand how TCP/IP shares network resources between users

\[
\text{maximize} \quad \sum_{i} u_i(x_i) \\
\text{subject to} \quad \sum_{i \in P(l)} x_i \leq c_l, \quad l \in L
\]
Lecture overview

Lecture 1: first-order methods for convex optimization

Lecture 2: decomposition techniques, application to multi-agent optimization

Part I:
Convex optimization using first-order methods

Aim: to understand
- properties and analysis techniques for basic gradient method
- the interplay between problem structure and convergence rate guarantees
- how we can deal with non-smoothness, noise and constraints
Rationale

Convex optimization:
- minimize convex function subject to convex constraints
- local minima global, strong and useful theory

First-order methods:
- only use function and gradient evaluations (i.e. no Hessians)
- easy to analyze, implement and distribute, yet competitive

Convex functions and convex sets

\[ \alpha x + (1 - \alpha) y \in X, \ \alpha \in [0, 1] \]
\[ \alpha f(x) + (1 - \alpha) f(y) \geq f(\alpha x + (1 - \alpha) y), \ \alpha \in [0, 1] \]
**Affine lower bounds from convexity**

Why?

\[ \alpha f(x) + (1 - \alpha)f(y) \geq f(\alpha x + (1 - \alpha)y) \Rightarrow \]

\[ f(y) \geq \frac{1}{1 - \alpha} \left( f(\alpha x + (1 - \alpha)y) - \alpha f(x) \right) = \]

\[ = f(x) + \frac{1}{1 - \alpha} \left( f(\alpha x + (1 - \alpha)y) - f(x) \right) \]

\[ = f(x) + \frac{1}{1 - \alpha} \left( f(x + (1 - \alpha)(y - x)) - f(x) \right) \]

Letting \( \alpha \to 0 \) yields result. Can also go in other direction

\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \]

---

**Strong convexity – quadratic lower bounds**

\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{c}{2} \| y - x \|^2 \]
Lipschitz continuous gradient – upper bounds

Lipschitz-continuous gradient: \( \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \)

Yields upper quadratic bound: 
\[
  f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| y - x \|^2
\]

Strongly convex functions with Lipschitz gradient

Bounded from above and below by quadratic functions

**Condition number** \( \kappa = \frac{L}{c} \) impacts performance of first-order methods.  
Note: limited function class when required to hold globally.
The basic gradient method

Basic gradient method
\[ x(t + 1) = x(t) - \alpha(t) \nabla f(x(t)) \]

A descent method (for small enough step-size \( \alpha(t) \)).

Convergence proof.
\[
\|x(t + 1) - x^*\|_2^2 = \|x(t) - x^*\|_2^2 - 2\alpha(t) \langle \nabla f(x(t)), x(t) - x^* \rangle + \alpha(t)^2 \| \nabla f(x(t)) \|_2^2 \\
\leq \|x(t) - x^*\|_2^2 - 2\alpha(t) (f(x(t)) - f^*) + \alpha(t)^2 \| \nabla f(x(t)) \|_2^2
\]

Where the inequality follows from convexity of \( f \)

---

Gradient method convergence proof

Applying recursively, we find
\[
\|x(T) - x^*\|_2^2 \leq \|x(0) - x^*\|_2^2 - 2 \sum_{t=0}^{T-1} \alpha(t) (f(x(t)) - f^*) + \sum_{t=0}^{T-1} \alpha^2(t) \| \nabla f(x(t)) \|_2^2
\]

Since gradient method is descent, and norms are non-negative
\[
2(f(x(T)) - f^*) \sum_{t=0}^{T-1} \alpha(t) \leq \|x(0) - x^*\|_2^2 + \sum_{t=0}^{T-1} \alpha^2(t) \| \nabla f(x(t)) \|_2^2
\]

Hence, with \( R_0 = \|x(0) - x^*\| \)
\[
f(x(T)) - f^* \leq \frac{R_0^2 + \sum_{t=0}^{T-1} \alpha^2(t) \| \nabla f(x(t)) \|_2^2}{2 \sum_{t=0}^{T-1} \alpha(t)}
\]

Further assumptions needed to guarantee convergence!
Gradient method discussion

If we assume that $f$ is Lipschitz, i.e. $\| \nabla f(x(t)) \| \leq L_f$

$$f(x(T)) - f^* \leq \frac{R_0^2 + L_f^2 \sum_{t=0}^{T-1} \alpha^2(t)}{2 \sum_{t=0}^{T-1} \alpha(t)}$$

Then,

- For fixed step-size $\alpha(t) = \alpha$
  $$\lim_{T \to \infty} f(x(T)) \leq f^* + \frac{\alpha L_f^2}{2}$$

- For diminishing stepsizes $\sum_{t=0}^{\infty} \alpha^2(t) < \infty$, $\sum_{t=0}^{\infty} \alpha(t) = \infty$
  $$\lim_{T \to \infty} f(x(T)) = f^*$$

- Accuracy $\varepsilon$ can be obtained in $(R_0 L_f)^2 / \varepsilon^2$ steps

Example

Smaller residual error for smaller stepsize, convergence for diminishing
**Strongly convex functions with Lipschitz gradient**

As in the basic gradient method proof

\[ \|x(t+1) - x^*\|^2_2 = \|x(t) - x^*\|^2_2 - 2\alpha(t)\langle \nabla f(x(t)), x(t) - x^* \rangle + \alpha^2(t)\|\nabla f(x(t))\|^2_2 \]

For strongly convex functions with Lipschitz-continuous gradient, it holds

\[ \langle \nabla f(x(t)), x(t) - x^* \rangle \geq \frac{cL}{c+L}\|x(t) - x^*\|^2_2 + \frac{1}{c+L}\|\nabla f(x(t))\|^2 \]

so

\[ \|x(t+1) - x^*\|^2_2 \leq \left(1 + \frac{2\alpha(t)cL}{c+L}\right)\|x(t) - x^*\|^2_2 + \alpha(t)\left(\frac{2}{c+L}\right)\|\nabla f(x(t))\|^2_2 \]

Hence, if \( \alpha(t) \leq \frac{2}{(c+L)} \) we obtain **linear convergence** rate

\[ \|x(t+1) - x^*\|^2_2 \leq \left(1 - \frac{2cL}{c+L}\alpha(t)\right)\|x(t) - x^*\|^2_2 \]

---

**Order-optimal methods**

The basic gradient method is **not** the optimal first-order method.
- optimal first-order methods typically use memory, e.g.

\[ x(t+1) = y(t) - L^{-1}\nabla f(y(t)) \]
\[ y(t+1) = x(t+1) + \frac{1 - \sqrt{R}}{1 + \sqrt{R}}(x(t+1) - x(t)) \]

Particularly useful when \( f \) is convex and has Lipschitz-continuous gradient
- from \( O(1/\varepsilon) \) to \( O(1/\sqrt{\varepsilon}) \)
- achieves optimal rate (same as basic gradient) also in other cases
- not always fastest first-order method
## Gradient methods: limits of performance

<table>
<thead>
<tr>
<th>Problem class</th>
<th>First-order method</th>
<th>Complexity</th>
<th>$\epsilon=1%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipschitz-continuous function</td>
<td>Gradient</td>
<td>$O(1/\epsilon^2)$</td>
<td>10,000</td>
</tr>
<tr>
<td>Lipschitz-continuous gradient</td>
<td>Gradient</td>
<td>$O(1/\epsilon)$</td>
<td>100</td>
</tr>
<tr>
<td>Optimal gradient</td>
<td>$O(1/\sqrt{\epsilon})$</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Strongly convex, Lipschitz gradient</td>
<td>Gradient</td>
<td>$\ln(1/\epsilon)$</td>
<td>2.3</td>
</tr>
<tr>
<td>Optimal gradient</td>
<td>$\ln(1/\epsilon)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Non-smooth convex functions: subgradients

Subgradient $s_x$ gives affine lower bound on convex function at $x$

$$f(y) \geq f(x) + \langle s_x, x - y \rangle$$

Subdifferential: set of all subgradients
The subgradient method

As the gradient method, but using subgradients instead

\[ x(t + 1) = x(t) - \alpha(t)s(t), \quad s(t) \in \partial f(x(t)) \]

**Not** a descent method.

Hence, cannot bound \( \sum_{t=0}^{T} \alpha(t)(f(x(t)) - f^*) \) as before. Rather, we find

\[ \inf_{t} f(x(t)) \leq f^* + R_0^2 + \frac{\sum_{t=0}^{T} \alpha^2(t)\|s(t)\|^2}{2\sum_{t=0}^{T} \alpha(t)} \]

If subgradients are bounded, then same conclusions as for gradient method. (step-size, convergence rates, ...)

Averages behave better...

The running averages of iterates

\[ \bar{x}(t) = \frac{1}{t} \sum_{k=0}^{t} x(k) \]

are often better-behaved than iterates themselves.

Specifically, if subgradients are bounded \( \|s_x\| \leq L \), then averages satisfy

\[ f(\bar{x}(T)) \leq f^* + \frac{\sqrt{2} R_0 L}{\sqrt{T}} \]

(note how “inf” is gone)
Gradient method for constrained optimization

Constrained minimization problem

\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in X
\end{align*}

If projections onto $X$ are easy to compute, can use \textit{projected gradient}

\[ x(t + 1) = P_X \{ x(t) - \alpha(t) \nabla f(x(t)) \} \]

Same convergence proof as before, since projections are non-expansive

\[ \| P_X \{ x \} - P_X \{ y \} \|^2 \leq \| x - y \|^2 \]

Beyond the basic methods

Smooth optimization of non-smooth functions
- epsilon-optimal solution to non-smooth problem requires many iterations
- often better to smooth function and apply order-optimal method

Exploiting structure
- when problem is smooth problem + easily-solvable non-smooth
- many current applications in compressed sensing, sparse optimization

...
Duality

Associated with every convex optimization problem

minimize $f_0(x)$
subject to $f_i(x) \leq 0$

$solve$ $Ax = b$

is an associated dual problem

maximize $g(\lambda, \mu)$
subject to $\lambda \geq 0$

where

$$g(\lambda, \mu) = \inf_x \left\{ f_0(x) + \sum_i \lambda_i f_i(x) + \mu^T (Ax - b) \right\}$$

**Advantage:** dual problem convex, has simple constraint set

Key properties of dual function

Dual function $g$ is always concave, may be non-smooth.

Dual function is a lower bound of optimal value when $\lambda \geq 0$

$$g(\lambda, \mu) = \inf_x f_0(x) + \sum_i \lambda_i f_i(x) + \mu^T (Ax - b) \leq f_0(x^*)$$

For convex problems, primal optimal value agrees with dual optimal value

$$g^* = \sup_{\lambda \geq 0, \mu} g(\lambda, \mu) = g(\lambda^*, \mu^*) = f_0(x^*)$$

e.g. when there is a feasible point satisfying inequality constraints strictly ("Slater condition")
Solving the dual problem

Dual function concave, but possibly non-smooth.

Dual problem often solved by projected (sub)-gradient method

\[
\begin{align*}
\lambda(t+1) &= P_+\{\lambda(t) + \alpha(t)s_\lambda(t)\} \quad s_\lambda \in \partial_\lambda g(\lambda(t), \mu(t)) \\
\mu(t+1) &= \mu(t+1) + \alpha(t)s_\mu(t) \quad s_\mu \in \partial_\mu g(\lambda(t), \mu(t))
\end{align*}
\]

Can do better when dual function is strongly concave, has Lipschitz gradient! (conditions for this will follow in next lecture...)

Summary of Lecture 1

First-order methods for convex optimization:
– gradient method: convergence proof and convergence rate estimates
– optimal methods: more states, but still only gradient information
– easy to implement, strong performance for certain problem classes

Non-smooth optimization
– subgradient method
– not a descent method, averaging gives better properties

Duality and the dual optimization problem
Part II: Decomposition techniques

Aim: to understand

- The basic idea of decomposition, coupling variables/constraints
- Dual decomposition: principle, advantages and challenges
- Application to multi-agent optimization

Basic idea of decomposition techniques

Decompose one complex problem into many small:

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in X
\end{align*}
\]
The trivial case

Separable objectives and constraints

\[
\begin{align*}
\text{minimize} & \quad \sum_{i} f_i(x_i) \\
\text{subject to} & \quad x_i \in X_i
\end{align*}
\]

Trivially separates into \( n \) decoupled subproblems

\[
\begin{align*}
\text{minimize} & \quad f_i(x_i) \\
\text{subject to} & \quad x_i \in X_i
\end{align*}
\]

that can be solved in parallel and combined.

The more interesting ones

Problems with \textbf{coupling constraints}

\[
\begin{align*}
\text{minimize} & \quad f_1(x_1) + f_2(x_2) \\
\text{subject to} & \quad x_1 + x_2 \leq c
\end{align*}
\]

Problems with \textbf{coupled objectives}

\[
\begin{align*}
\text{minimize} & \quad f_1(x_1, x_{12}) + f_2(x_{12}, x_2)
\end{align*}
\]

Coupled objectives can be cast as a problem of coupling constraints:

\[
\begin{align*}
\text{minimize} & \quad f_1(x_1, z_{12}) + f_2(z_{21}, x_2) \\
\text{subject to} & \quad z_{12} = z_{21}
\end{align*}
\]

so this case will be our focus.
**Dual decomposition**

Basic idea: decouple problem by relaxing coupling constraints.

\[
\begin{align*}
\text{minimize} & \quad f_1(x_1) + f_2(x_2) \\
\text{subject to} & \quad x_1 + x_2 \leq c
\end{align*}
\]

Formally, introduce Lagrange multiplier for the constraint, form Lagrangian

\[L(x, \lambda) = f_1(x_1) + f_2(x_2) + \lambda(x_1 + x_2 - c)\]

with associated dual function

\[g(\lambda) = \inf_x L(x, \lambda) = -\lambda c + \inf_{x_1} \{f_1(x_1) + \lambda x_1\} + \inf_{x_2} \{f_2(x_2) + \lambda x_2\}\]

and solve the dual problem.

**Dual decomposition cont’d**

Dual problem has the form

\[
\begin{align*}
\text{maximize} & \quad g_1(\lambda) + g_2(\lambda) \\
\text{subject to} & \quad \lambda \geq 0
\end{align*}
\]

additive (hence, can be evaluated in parallel) and simple constraints.

The dual function is always concave, and a subgradient of \(g\) is given by

\[x_1^*(\lambda) + x_2^*(\lambda) - c\]

Hence, dual problem is convex. Can solve using projected subgradient method.
Dual and distributed optimization

Dual decomposition often results in additive dual function

\[ g_1(\lambda) \]
\[ \vdots \]
\[ g_n(\lambda) \]
\[ \sum \]
\[ g(\lambda) \]

but might still need coordinator to solve dual optimization problem.

Dual problem fully distributed if (sub)gradient of dual locally available

Hycon2 Workshop, ECC 2013
Mikael Johansson mikaelj@ee.kth.se

---

Dual decomposition example

Simple example:

\[
\begin{align*}
\text{minimize} & \quad |x_1 - 1| + |x_2 - 1| \\
\text{subject to} & \quad x_1 + x_2 \leq 1 \\
& \quad x_1 \in [-10, 10]
\end{align*}
\]

Optimal value \( f^*_0 = 1 \) for \( x_1^* = 1 - x_2^*, \quad x_2^* \in [0, 1] \)
Drawback of dual decomposition

Optimizes dual variables, to find optimal value of dual function.

\[
\begin{align*}
\text{maximize} & \quad g(\lambda) \\
\text{subject to} & \quad \lambda \geq 0 \quad \Rightarrow \lambda^* \text{, } d^* = g(\lambda^*)
\end{align*}
\]

In general, primal iterates might be suboptimal, violate constraints.

\[x^*(\lambda) = \arg \inf_x L(x, \lambda)\]

Under strong convexity of primal, and the existence of a Slater point:
- feasibility and primal optimality recovered in the limit.

Constraints and demands on subsystem consistency should be “soft”

Primal convergence in dual methods

Several techniques for enforcing primal convergence, e.g. averaging iterates

\[\bar{x}^*(t) = \frac{1}{t} \sum_{k=0}^{t} x^*(\lambda(t))\]

Under Slater, iterate average satisfies constraints asymptotically and

\[f_0(\bar{x}^*(t)) \leq f^* + \frac{\alpha \Delta^2}{2} + \frac{\|\lambda(0)\|^2}{2\alpha t}\]
Example

Simple example from before. Iterates and running averages:

---

Primal convergence in dual methods

Stronger properties when dual function is differentiable, strongly concave.

**Fact.** Consider the linearly constrained convex optimization problem

\[
\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad Ax = b \\
& \quad Cx \preceq d
\end{align*}
\]

If objective is strongly convex w. modulus \( c \), has Lipschitz-continuous gradient, and there exists a Slater point. Then, if

\[
A = [A^T \ C^T]^T
\]

has full row rank, iterates \( u = (\lambda, \mu) \) produced by dual projected gradient satisfy

\[
\|u(t) - u^*\| \leq q^t\|u(0) - u^*\|
\]

\[
\|x^*(u(t)) - x^*\| \leq q^t \frac{\sigma_{\max}(A)}{c}\|u(0) - u^*\|
\]

---
Application to multi-agent optimization

A network of agents collaborate to solve the optimization problem

\[
\text{minimize } \sum_{i \in V} f_i(x)
\]

Agents can only exchange information with neighbors in graph \(G = (V, E)\)

\[
f_1(x) \quad f_2(x) \quad f_3(x)
\]

1 \quad 2 \quad 3

Three techniques in some detail:

– dual decomposition, consensus-gradient, alternating direction of multipliers

Method 1: dual decomposition

Introduce local copy \(x_i\) of decision variable, re-write problem on the form

\[
\text{minimize } \sum_{i \in V} f_i(x_i)
\]

subject to

\[x_i = x_j \quad \forall (i, j) \in E\]

Relax consistency constraints using Lagrange multipliers, solve dual problem.
The dual decomposition approach

Convenient to write problem as

\[
\begin{align*}
\text{minimize} & \quad \sum_{i \in V} f_i(x_i) \\
\text{subject to} & \quad Mx = 0
\end{align*}
\]

where \( M \) is the edge-node incidence matrix of \( G \),

\[
[M]_{e,i} = \begin{cases} 
1 & \text{if } i \text{ is the start node of edge } e \\
-1 & \text{if } i \text{ is the end node of edge } e \\
0 & \text{otherwise}
\end{cases}
\]

Introducing Lagrange multiplier vector \( \mu \in \mathbb{R}^{|E|} \), form Lagrangian

\[
L(x, \mu) = \sum_{i \in V} f_i(x_i) + \mu^T Mx = \sum_{i \in V} f_i(x_i) + \sum_{j: (i,j) \in E} \mu_{ij}(x_i - x_j)
\]

Dual decomposition updates become

\[
x_i(t+1) = \arg\min_{x_i} L(x_i, \mu) = \arg\min_{x_i} \left( f_i(x_i) + \sum_{j: (i,j) \in E} \mu_{ij}(t)x_i - \sum_{j: (j,i) \in E} \mu_{ji}(t)x_i \right)
\]

\[
\mu_{ij}(t+1) = \mu_{ij}(t) + \alpha(t) (x_i(t+1) - x_j(t+1))
\]

Data exchange only between neighbors.

Does iterations converge? Under what assumptions? Good stepsizes?
Method 2: consensus-gradients

Use same modeling idea, i.e. consider

\[
\begin{align*}
& \text{minimize} & & \sum_{i \in V} f_i(x_i) \\
& \text{subject to} & & Mx = 0 \\
\end{align*}
\]

Replace strict equalities with penalty term

\[
\text{minimize } p(x) := \sum_{i \in V} f_i(x_i) + \frac{\eta}{2} \|Mx\|^2
\]

Note: an optimality-consistency trade-off

Gradient descent on penalty function

The gradient iterations become

\[
x(t + 1) = x(t) - \alpha(t) \frac{\partial}{\partial x_i} p(x) = x(t) - \alpha(t) (\nabla f(x(t)) + \eta M^T M x)
\]

which we can re-write as

\[
x_i(t + 1) = x_i(t) + \sum_{j : (i,j) \in \mathcal{E}} \alpha(t) \eta (x_j(t) - x_i(t)) - \alpha(t) \nabla f_i(x_i(t))
\]

“consensus”

A combination of fixed-weight consensus and gradient descent.
**Consensus-subgradient method**

Originally proposed for non-smooth optimization

\[ x_i(t + 1) = \left\{ W_{ii}x_i(t) + \sum_{j:(i,j) \in \mathcal{E}} W_{ij}x_j(t) \right\} - \alpha_i s(t), \quad s(t) \in \partial f(x(t)) \]

Studied under general consensus weights, time-varying graphs.

For fixed step-sizes, iterations do not converge to true optimum
- need average iterates, use diminishing stepsizes

**Method 3: ADMM**

Alternating direction of multipliers (ADMM) considers problem on the form

minimize \[ f(x) + g(z) \] subject to \[ Ex + Fz = h \]

\[ \iff \]

minimize \[ f(x) + g(z) + \frac{\mu}{2}||Ex + Fz - h||_2^2 \] subject to \[ Ex + Fz = h \]

Finds optimal solution by alternating minimization of augmented Lagrangian

\[ L_\mu(x, z, \mu) = f(x) + g(z) + \mu^T(Ex + Fz - h) + \frac{\mu}{2}||Ex + Fz - h||_2^2 \]

followed by Lagrange multiplier update, i.e.:

\[ x(t + 1) = \arg\min_x L_\mu(x, z(t), \mu(t)) \]

\[ z(t + 1) = \arg\min_z L_\mu(x(t + 1), z, \mu(t)) \]

\[ \mu(t + 1) = \mu(t) + \rho(Ex(t + 1) + Fz(t + 1) - h) \]
ADMM properties

Under mild conditions, ADMM converges for all values of $\rho > 0$ (in contrast to dual methods, where large step-size can cause divergence).

Convergence rates of ADMM is a topic of intense current research.

The penalty parameter $\rho$ affects the convergence factors of the iterates.

– optimal parameter selection rules exist for some problem classes

ADMM for quadratic problems

Quadratic programming problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} x^T P x + q^T x \\
\text{subject to} & \quad A x \leq b
\end{align*}
\]

re-written on ADMM standard form

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} x^T P x + q^T x + \mathcal{I}_+(z) \\
\text{subject to} & \quad A x - b + z = 0
\end{align*}
\]

yields iterations

\[
\begin{align*}
x(t + 1) &= -(Q + \rho A^T A)^{-1} [q + \rho A^T (z(t) + u(t) - b)] \\
z(t + 1) &= \max\{0, -A(t+1) - u(t) + b\} \\
u(t + 1) &= u(t) + Ax(t + 1) - b + z(t + 1)
\end{align*}
\]

What can we say about convergence, optimal $\rho$?
ADMM for quadratic problems

Fact. For all $\rho > 0$ ADMM iterations converge to optimum at linear rate.

Fact. If $A$ is invertible or has full row rank, then

$$\rho = \frac{1}{\sqrt{\lambda_1(AQ^{-1}A^T)\lambda_n(AQ^{-1}A^T)}}$$

yields the smallest convergence factor (fastest convergence times).

(tends to work well also when $A$ does not have these properties)

ADMM for multi-agent optimization

Introduce “agreement variable” $z_{(i,j)}$ on each edge $(i, j) \in \mathcal{E}$, consider

minimize $\sum_{i \in V} f_i(x_i)$
subject to $x_i = z_{(i,j)} \quad \forall (i, j) \in \mathcal{E}$
$z_{(i,j)} = z_{(i,j)} \quad \forall (i, j) \in \mathcal{E}$

Can be re-written as

minimize $\sum_{i \in V} f_i(x_i)$
subject to $\begin{bmatrix} M_+ \\ M_- \end{bmatrix} \begin{bmatrix} x \\ I \end{bmatrix} - \begin{bmatrix} I \\ I \end{bmatrix} z = 0$

where $M_+ = \max\{M, 0\}$, $M_- = -\min\{M, 0\}$
ADMM for multi-agent optimization

ADMM iterations become

\[
x_i(t + 1) = \arg\min_x f_i(x) + (\mu_{ij} + \mu_{ji})x + \frac{\rho}{2} (x - z_{ij})^2 + (x - z_{ji})^2
\]

\[
z_{ij}(t + 1) = \rho x_i(t + 1) + \mu_{ij}(t)
\]

\[
\mu_{ij}(t + 1) = \mu_{ij}(t) + \rho(x_i(t + 1) - z_{ij}(t + 1))
\]

Converge for all values of penalty parameter.

Many variations, extensions (e.g. different penalty parameters per edge)

Example: robust estimation

Nodes measure different noisy versions \(y_i(t)\) of the same quantity.

Would like to agree on common estimate \(\hat{x}\) that minimizes

\[
\text{minimize} \quad \sum_{i \in V} \|y_i - x\|_H
\]

subject to \(x \in X\)

\(G = (V, \mathcal{E})\)

where \(\| \cdot \|_H\) is the Huber loss
Example: robust optimization

Representative results, 100-node ring network

Summary of Lecture 2

Dual decomposition: idea and properties.

Multi-agent optimization:
  – collaborative optimization under information exchange constraints

Three techniques in (some) detail
  – Dual decomposition
  – ADMM
  – Gradient/consensus method

Many alternative techniques not covered.
So what did we see?

Lecture 1: first-order methods for convex optimization

Lecture 2: dual decomposition and optimization over graphs

References for Lecture 1

Lecture one is covered, at least in parts, in many textbooks. The books

B. Polyak, “Introduction to optimization”, 1987
Y. Nesterov, “Introductory lectures on convex optimization: a basic course”, 2004

are particularly beautiful accounts. A good reference for duality theory is

D. Bertsekas, A. Ozdaglar, A. Nedich, “Convex analysis and optimization"
References for Lecture 2

The material on dual decomposition is based on the chapter
B. Yang and M. Johansson, “Distributed optimization, a tutorial overview”
from the “Networked Control” book of an earlier Hycon Summer School. The book covers many individual references to original work by a wide range of authors.

The lecture notes

has a nice introduction to modelling for distributed optimization.

References for lecture 2

The survey paper
S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers”, 2010
covers theory and applications of ADDM. Optimal penalty parameter selection is studied in

Subgradient-consensus techniques were proposed in