Primal Solutions and Rate Analysis for Subgradient Methods

Asu Ozdaglar

Joint work with Angelia Nedić, UIUC

Conference on Information Sciences and Systems (CISS)
March, 2008

Department of Electrical Engineering & Computer Science

Massachusetts Institute of Technology

Introduction

- Lagrangian relaxation and duality effective tools for
  - solving large-scale convex optimization,
  - systematically providing lower bounds on the optimal value
- Subgradient methods provide efficient computational means to solve the dual problem to obtain
  - Near-optimal dual solutions
  - Bounds on the primal optimal value
- Most remarkably, in networking applications, subgradient methods have been used to design **decentralized resource allocation mechanisms**
  - Kelly 1997, Low and Lapsley 1999, Srikant 2003, Chiang *et al.* 2007

Issues with this approach

- Subgradient methods **operate in the dual space**
  - In most problems, interest in primal solutions
- Convergence analysis mostly focuses on diminishing stepsize
- No convergence rate analysis
- **Question of Interest:** Can we use the subgradient information to produce near-feasible and near-optimal primal solutions?

Our Work

- Primal solution generation from subgradient algorithms
- Main Results:
  - Development of algorithms that use the subgradient information and an **averaging scheme** to generate approximate primal optimal solutions
  - Convergence rate analysis for the approximation error of the primal solutions including:
    - The amount of feasibility violation
    - Primal optimal cost approximation error
  - Stopping criteria for our algorithms
- This talk has two parts:
  - Dual subgradient algorithms (subgradient of the dual function available)
  - Primal-dual subgradient algorithms

Prior Work

- Subgradient methods producing primal solutions by averaging
  - Nemirovskii and Yudin 1978
  - Shor 1985, Sherali and Choi 1996 [linear primal]
  - Larsson, Patriksson, Strömberg 1995, 1998, 1999 [convex primal]
  - Kiwiel, Larsson, and Lindberg 2007
- In all of the existing literature:
  - Interest is in generating primal optimal solutions **in the limit**
  - The focus is on subgradient algorithms using a diminishing step
  - There is no convergence rate analysis
- (Primal) subgradient methods that use averaging to generate solutions
  - Nesterov 2005, Ruszczynski 2007

Primal and Dual Problem

- We consider the following **primal problem**

$$
\begin{aligned}
f^* = \text{minimize} \quad & f(x) \\
\text{subject to} \quad & g(x) \le 0,\ x \in X,
\end{aligned}
$$

  where $g(x) = (g_1(x), \ldots, g_m(x))$ and f^* is finite.

  – The functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ are convex, and the set $X \subseteq \mathbb{R}^n$ is nonempty and convex

- We are interested in solving the primal problem by considering the **Lagrangian dual problem**

$$
\begin{aligned}
q^* = \text{maximize} \quad & q(\mu) = \inf_{x \in X} \{f(x) + \mu^T g(x)\} \\
\text{subject to} \quad & \mu \ge 0,\ \mu \in \mathbb{R}^m
\end{aligned}
$$

Dual Subgradient Method

The dual iterates are generated by the following update rule:

$$\mu_{k+1} = [\mu_k + \alpha_k g_k]^+ \quad \text{for } k \geq 0$$

- μ_0 is an initial iterate with $\mu_0 \geq 0$
- $[\cdot]^+$ denotes the projection on the nonnegative orthant
- $\alpha_k > 0$ is a stepsize
- g_k is a subgradient of $q(\mu)$ at μ_k, i.e.,

$$g_k = g(x_k) \quad \text{with} \quad x_k \in X \quad \text{and} \quad q(\mu_k) = f(x_k) + \mu_k^T g(x_k)$$

We assume that:

- The set of optimal solutions, $\arg\min_{x \in X}\{f(x) + \mu^T g(x)\}$, is nonempty for all $\mu \geq 0$
- The subgradient of the dual function is "easy" to compute

Dual Set Boundedness under Slater

Assumption *(Slater Condition)* There is a vector $\bar{x} \in \mathbb{R}^n$ such that

$$g_j(\bar{x}) < 0, \qquad \forall\ j = 1, \ldots, r.$$

Under the Slater condition, we have:

- The dual optimal set is nonempty and **bounded**
- There holds for any dual optimal solution $\mu^* \geq 0$,

$$\sum_{j=1}^{m} \mu_j^* \leq \frac{f(\bar{x}) - q^*}{\min_{1 \leq j \leq m}\{-g_j(\bar{x})\}} \qquad \text{[Uzawa 1958]}$$

We extend this result, as follows:

Proposition: Let the Slater condition hold. Then, for every $c \in \mathbb{R}$, the set $Q_c = \{\mu \geq 0 \mid q(\mu) \geq c\}$ is bounded:

$$\|\mu\| \leq \frac{f(\bar{x}) - c}{\min_{1 \leq j \leq m}\{-g_j(\bar{x})\}} \qquad \text{for all } \mu \in Q_c$$

where $\bar{x}$ is a Slater vector.

Analysis of the Subgradient Method

Consider the algorithm with a constant stepsize $\alpha > 0$, i.e.,

$$\mu_{k+1} = [\mu_k + \alpha g_k]^+ \quad \text{for } k \geq 0$$

Assumption *(Bounded Subgradients)* The subgradient sequence $\{g_k\}$ is bounded, i.e., there exists a scalar $L > 0$ such that

$$\|g_k\| \leq L, \qquad \forall \ k \geq 0$$

- This assumption satisfied when primal constraint set X is compact
  - By the convexity of the g_j over $\mathbb{R}^n$, $\max_{x \in X} \|g(x)\|$ is finite and provides an upper bound on the norms of the subgradients

Bounded Multipliers

Proposition: Let the Slater condition hold and let the subgradients g_k be bounded. Let $\{\mu_k\}$ be the multiplier sequence generated by the subgradient algorithm. Then, the sequence $\{\mu_k\}$ is bounded. In particular, for all k, we have

$$\|\mu_k\| \leq \frac{2}{\gamma}\left[f(\bar{x}) - q^*\right] + \max\left\{\|\mu_0\|,\ \frac{1}{\gamma}\left[f(\bar{x}) - q^*\right] + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}$$

- α is the stepsize
- $\bar{x}$ is a Slater vector
- $\gamma = \min_{1 \leq j \leq m}\{-g_j(\bar{x})\}$
- L is a subgradient norm bound

Subgradient Algorithm and Primal Averages

Subgradient Method

Generates multipliers in the dual space:

$$\mu_{k+1} = [\mu_k + \alpha g_k]^+ \qquad \text{for } k \geq 0$$

$$g_k = g(x_k) \qquad \text{with } x_k \in X \text{ and } q(\mu_k) = f(x_k) + \mu_k^T g(x_k)$$

Primal Averaging

Generates the primal averages of $x_0, \ldots, x_{k-1}$:

$$\hat{x}_k = \frac{1}{k} \sum_{i=0}^{k-1} x_i \qquad \text{for } k \geq 1$$

- Each $\hat{x}_k$ belongs to X by convexity of X and the fact $x_i \in X$ for all i
- The vectors $\hat{x}_k$ need not be feasible
- We consider $\hat{x}_k$ as an *approximate primal solution*

Basic Estimates for the Primal Averages

Proposition:

Let $\{\mu_k\}$ be generated by the subgradient method with a stepsize α.

Let $\hat{x}_k$ be the primal averages of the subgradient defining vectors $x_k \in X$.

Then, for all $k \geq 1$:

- The amount of feasibility violation at $\hat{x}_k$ is bounded by

$$\left\| g(\hat{x}_k)^+ \right\| \leq \frac{\|\mu_k\|}{k\alpha}$$

- The primal cost at $\hat{x}_k$ is bounded above by

$$f(\hat{x}_k) \leq q^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha}{2k} \sum_{i=0}^{k-1} \|g(x_i)\|^2$$

- The primal cost at $\hat{x}_k$ is bounded below by

$$f(\hat{x}_k) \geq q^* - \|\mu^*\| \, \left\| g(\hat{x}_k)^+ \right\|$$

  where μ^* is a dual optimal solution and q^* is the dual optimal value.

Estimates under Slater

Proposition: Let Slater condition hold and subgradients be bounded. Then, the estimates for $\hat{x}_k$ can be strengthened as follows: for all $k \geq 1$,

- The amount of feasibility violation is bounded by

$$\left\| g(\hat{x}_k)^+ \right\| \leq \frac{B^*_{\mu_0}}{k\alpha}$$

- The primal cost is bounded above by

$$f(\hat{x}_k) \leq f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2}$$

- The primal cost is bounded below by

$$f(\hat{x}_k) \geq f^* - \frac{1}{\gamma} \left[f(\bar{x}) - q^* \right] \left\| g(\hat{x}_k)^+ \right\|$$

where L is a subgradient norm bound, $\gamma = \min_{1 \leq j \leq m} \{-g_j(\bar{x})\}$

$$B^*_{\mu_0} = \frac{2}{\gamma} \left[f(\bar{x}) - q^* \right] + \max \left\{ \|\mu_0\|, \ \frac{1}{\gamma} \left[f(\bar{x}) - q^* \right] + \frac{\alpha L^2}{2\gamma} + \alpha L \right\}$$

Analyzing the Results

Choosing $\mu_0 = 0$ yields:

$$\left\|g(\hat{x}_k)^+\right\| \leq \frac{B_0^*}{k\alpha} \qquad \text{with} \quad B_0^* = \frac{3}{\gamma}\left[f(\bar{x}) - q^*\right] + \frac{\alpha L^2}{2\gamma} + \alpha L$$

$$f(\hat{x}_k) \leq f^* + \frac{\alpha L^2}{2}$$

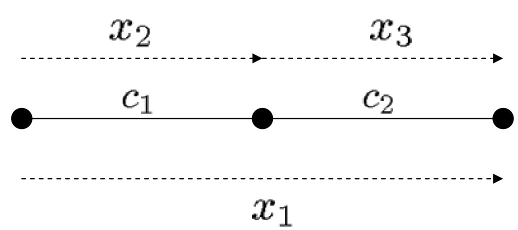
$$f(\hat{x}_k) \geq f^* - \frac{1}{\gamma}\left[f(\bar{x}) - q^*\right] \left\|g(\hat{x}_k)^+\right\|$$

Remarks:

- The rate of convergence to the primal "near-optimal" value is driven by the rate of infeasibility decrease
- The bound on feasibility violation B_0^* involves dual optimal value q^*. We can use $\max_{0 \leq i \leq k} q(\mu_i) \leq q^*$ for an alternative bound.
- **Stopping criteria readily available** from these estimates
- The estimates capture the trade-offs between a desired accuracy and the computations required to achieve the accuracy

Example

- Rate allocation using **network utility maximization**
- A simple network with 2 serial links and 3 sources, with rate x_i
- Link capacities are $c_1 = 1$ and $c_2 = 2$
- Each user has utility function $u_i(x_i) = \sqrt{x_i}$



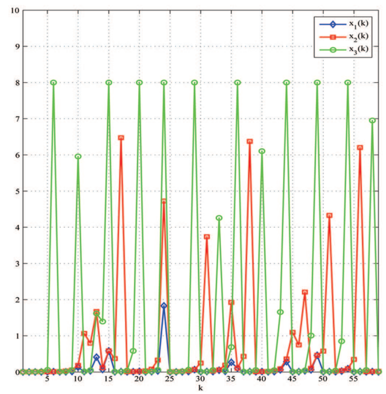
- We allocate rates as the optimal solution of

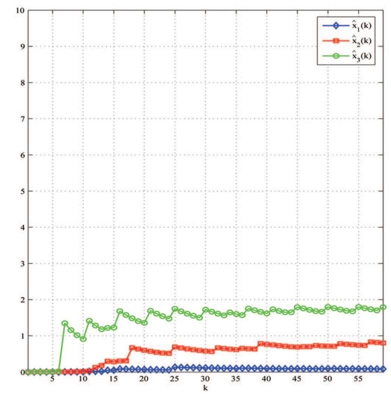
$$
\begin{array}{ll}
\text{maximize} & \sum_{i=1}^{3} \sqrt{x_i} \\
\text{subject to} & x_1 + x_2 \leq 1, \quad x_1 + x_3 \leq 2, \\
& x_i \geq 0, \quad i = 1, 2, 3.
\end{array}
$$

- We use a dual subgradient method and averaging to generate primal solutions

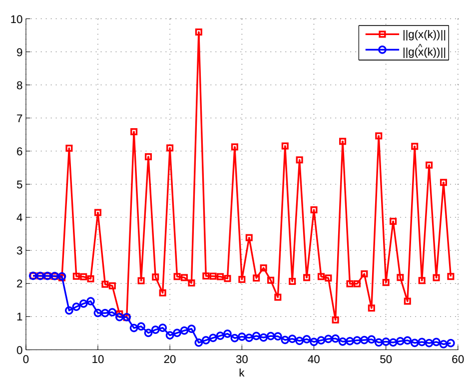
Performance

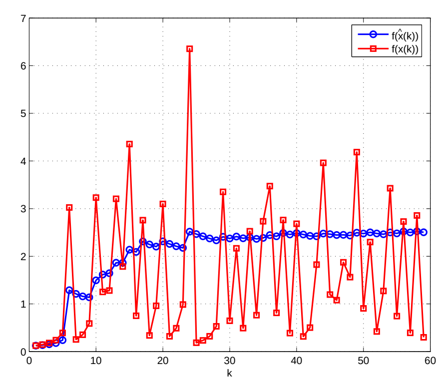
- The convergence behavior of the primal sequence $\{x_k\}$ (left) and $\{\hat{x}_k\}$ (right)





- The convergence behavior of the constraint violation (left) and primal objective function values (right) for the two primal sequences





Primal-Dual Subgradient Method

- Assume subgradient of dual function cannot be computed efficiently
- We consider methods for computing saddle point of Lagrangian

$$\mathcal{L}(x, \mu) = f(x) + \mu' g(x), \qquad \text{for all } x \in X, \ \mu \geq 0$$

Primal-Dual Subgradient Method:

$$x_{k+1} = \mathcal{P}_X \left[x_k - \alpha \mathcal{L}_x(x_k, \mu_k) \right] \qquad \text{for } k = 0, 1, \ldots.$$

$$\mu_{k+1} = \mathcal{P}_D [\mu_k + \alpha \mathcal{L}_\mu(x_k, \mu_k)] \qquad \text{for } k = 0, 1, \ldots.$$

- D is a closed convex set containing set of dual optimal solutions
- $\mathcal{L}_x(x_k, \mu)$ denotes a subgradient wrt x of $\mathcal{L}(x, \mu)$ at x_k.
- $\mathcal{L}_\mu(x, \mu_k)$ denotes a subgradient wrt μ of $\mathcal{L}(x, \mu)$ at μ_k.

$$\mathcal{L}_x(x_k, \mu) = s_f(x_k) + \sum_{i=1}^{m} \mu_i s_{g_i}(x_k), \qquad \mathcal{L}_\mu(x, \mu_k) = g(x),$$

where $s_f(x_k)$ and $s_{g_i}(x_k)$ are subgradients of f and g_i at x_k.

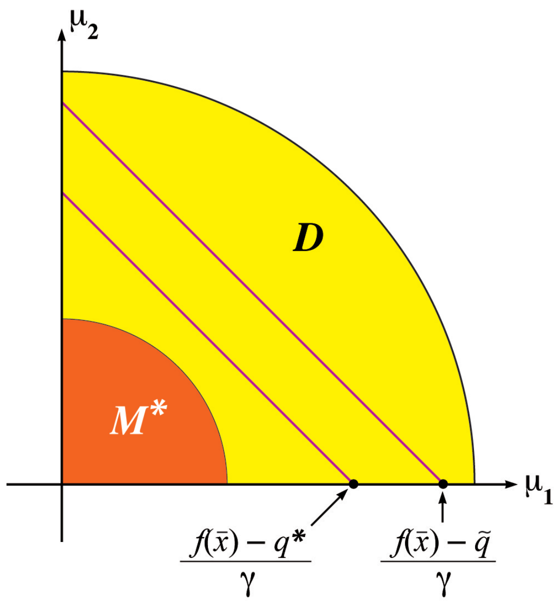
- Builds on the seminal Arrow-Hurwicz-Uzawa gradient method 1958

Set D under Slater Assumption

- Under Slater, dual optimal set M^* nonempty and bounded
- This motivates the following choice for set D:

$$D = \left\{ \mu \geq 0 \mid \|\mu\| \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r \right\}$$

where $r > 0$ is a scalar parameter



Assumption (*Compactness*): Set X is compact, $\|x\| \leq B$, for all $x \in X$.

- Under the assumptions and the definition of the method, the subgradients are bounded:

$$\max_{k \geq 0} \max \left\{ \|\mathcal{L}_x(x_k, \mu_k)\|, \|\mathcal{L}_\mu(x_k, \mu_k)\| \right\} \leq L.$$

- The subgradient boundedness was assumed in previous analysis (Gol'shtein 72, Korpelevich 76)

Estimates for the Primal-Dual Method

Proposition: Let the Slater and Compactness Assumptions hold. Let $\{\hat{x}_k\}$ be the primal average sequence. Then, for all $k \geq 1$, we have:

- The amount of feasibility violation is bounded by

$$\|g(\hat{x}_k)^+\| \leq \frac{2}{k\alpha r} \left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} + r \right)^2 + \frac{\|x_0 - x^*\|^2}{2k\alpha r} + \frac{\alpha L^2}{2r}.$$

- The primal cost is bounded above by

$$f(\hat{x}_k) \leq f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\|x_0 - x^*\|^2}{2k\alpha} + \alpha L^2.$$

- The primal cost is bounded below by

$$f(\hat{x}_k) \geq f^* - \left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} \right) \|g(\hat{x}_k)^+\|.$$

Optimal Choice for r and Resulting Estimate

By minimizing the bound for the feasibility violation with respect to the parameter $r > 0$, we obtain:

- The resulting optimal r^* depends on the iteration index k:

$$r^*(k) = \sqrt{\left(\frac{f(\bar{x}) - \tilde{q}}{\gamma}\right)^2 + \frac{\|x_0 - x^*\|^2}{4} + \frac{k\alpha^2 L^2}{4}} \qquad \text{for } k \geq 1.$$

Given some k, consider an algorithm where dual iterates are obtained by

$$\mu_{i+1} = \mathcal{P}_{D_k}[\mu_i + \alpha \mathcal{L}_\mu(x_i, \mu_i)], \qquad D_k = \left\{ \mu \geq 0 \,\middle|\, \|\mu\| \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r^*(k) \right\}$$

- The resulting feasibility violation estimate at the primal average $\hat{x}_k$:

$$\|g(\hat{x}_k)^+\| \leq \frac{8}{k\alpha}\left(\frac{f(\bar{x}) - \tilde{q}}{\gamma}\right) + \frac{2\|x_0 - x^*\|}{k\alpha} + \frac{2L}{\sqrt{k}}$$

Conclusions

- We considered dual and primal-dual subgradient methods with primal averaging to generate primal "near-feasible" and "near-optimal" solutions
- Slater assumption plays a key role in our analysis
- We provided estimates for feasibility violation and primal cost
- Our estimates capture the trade-offs between desired accuracy and the computations required to achieve the accuracy
- Our analysis shows that
  - The scheme using dual subgradient method converges with rate $1/k$
  - The scheme using primal-dual subgradient method converges with rate $1/\sqrt{k}$