Lecture Notes for
Math 582, Winter 2009
Univ. Washington
on

Convex Optimization Algorithms

Week 1

Paul Tseng[1]

[1]Department of Mathematics, University of Washington, Seattle, WA 98195, U.S.A. (tseng@math.washington.edu)

1 What?

Convex optimization problems arising from applications or as approximations of intractable problems are often large, but structured. Exploiting the structures, possibility through duality, is key to solving these problems efficiently. We will examine different types of structures and, for each, look for algorithms that best exploit the structures. The problem types include quadratic, conic (semidefinite cone, second-order cone), smooth, and "simple" non-smooth. The algorithms include first-order gradient methods and second-order Newton methods (e.g., interior-point methods). Relevant issues include approximation bounds, convergence, complexity, and implementation.

Below is a list of topics that we hope to cover:

- Motivation
- Background on convex analysis, covex optimization, duality
- Interior-point methods
- Gradient methods
- Incremental, coordinate gradient methods, simplicial decomposition (maybe?)
- Approximation bounds

These notes are supposed to be integrated into a book that I am writing with Dimitri Bertsekas, so any questions/comments and (gentle) criticisms are welcome! Below is a list of references (annotated).

References

[1] Ben-Tal, A. and Nemirovski, A., Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications, SIAM, Philadelphia, 2001. (Well written and has a good treatment of conic optimization, interior-point methods, and applications. The second author, in particular, has made important contributions to interior-point methods.)

[2] Bertsekas, D. P., Nonlinear Programming, 2nd edition, Athena Scientific, Belmont, 1999. (Well written and has a good treatment of gradient methods and dual methods.)

[3] Boyd, S. and Vandenberghe, L., Convex Optimization, Cambridge University Press, Cambridge, 2004. (Well written and has many examples and applications, with some treatment of algorithms.)

[4] Boyd, S., El Ghaoui, L., Feron, E., and Blakrishnan V., Linear Matrix Inequalities in System and Control Theory, SIAM, Philadelphia, 1994. (A comprehensive monograph on LMI.)

[5] Goemans, M. X., Semidefinite programming and combinatorial optimization, Proc. Intern. Congress Math., Vol. III (Berlin, 1998), 657–666. (A bit old, but still good survey of approximating intractable problems by SDP.)

[6] Klerk, E. de, Aspects of Semidefinite Programming: Interior Point Algorithms and Selected Applications, Kluwer, Dordrecht, 2002. (I have not read it, but it could be worth a look.)

[7] Hiriart-Urruty, J.-B. and Lemaréchal, C., Convex Analysis and Minimization Algorithms, Vol. I and II, Springer, Berlin, 1993. (Somewhat old and specialized.)

[8] Nesterov, Y., Introductory Lectures on Convex Optimization, Kluwer Academic Publisher, Dordrecht, The Netherlands, 2004. (Has some recent results on interior-point methods and gradient methods. It's between a book and a monograph, and much easier to understand than [9].)

[9] Nesterov, Y. and Nemirovskii, A., Interior Point Polynomial Methods in Convex Programming: Theory and Applications, SIAM, Philadelphia, 1993. (A monograph on interior-point method for convex optimization. It has very original ideas, especially on self-concordant functions, but is hard to understand.)

[10] Rockafellar, R. T., Convex Analysis, Princeton Univ. Press, Princeton, 1970. (A classic. Excellent as a reference, but not easy to learn from.)

[11] Wolkowicz, H., Saigal, R., and Vandenberghe, L., editors, Handbook of Semidefinite Programming, Kluwer, Boston, 2000. (A collection of survey articles on SDP.)

[12] Wright, S. J., Primal-Dual Interior-Point Methods, SIAM, Philadelphia, 1997. (For LP, this seems the most popular reference on interior-point method.)

[13] Ye, Y., Interior Point Algorithms: Theory and Analysis, John Wiley & Sons, New York, 1997. (I have not read it, but it's worth a look as the author has made important contributions to interior-point methods.)

2 Motivation

Convex optimization arises in many guises, both in applications and as approximation to intractable problems; see [1, 3, 4, 5]. Let's look at some recent examples. We'll see other examples as the course moves along.

2.1 Compressed sensing

A problem in signal processing that has received much attention is that of compressed sensing. In the basic version of this problem, we wish to find a sparse representation of a given (discretized)

noiseless signal $b \in \mathbb{R}^m$ from a dictionary of n elementary signals (e.g., Wavelets), guided by Occam's Razor principle that "simplest is best." This may be formulated as

$$\begin{array}{ll} \min_x & \sharp(x) \\ \text{s.t.} & Ax = b \end{array} \tag{1}$$

where $A \in \mathbb{R}^{m \times n}$ comprises the elementary signals for its columns and $\sharp(x)$ counts the number of nonzero components in $x \in \mathbb{R}^n$. In typical applications, m and n are large ($m, n \geq 2000$). This problem is known to be intractable (NP-hard).[2] (We won't go into detail on what is an NP-hard problem. Roughly speaking, it's a problem for which no "polynomial-time" algorithm is known. Moreover, *if* a polynomial-time algorithm can be found, then P = NP, i.e., all problems in the class NP are solvable in polynomial time.)

A popular solution approach is to approximate it by a convex problem, with $\sharp(\cdot)$ replaced by the 1-norm $\|\cdot\|_1$ (i.e., $\|x\|_1 = \sum_{j=1}^n |x_j|$), which is a convex function; also see [3, Section 6.2]. This results in

$$\begin{array}{ll} \min_x & \|x\|_1 \\ \text{s.t.} & Ax = b \end{array} \tag{2}$$

which is a convex problem. In fact it can be transformed into a linear program (LP) by making the substitution $|x_j| = u_j + v_j$, with $x_j = u_j - v_j$ and $u_j \geq 0, v_j \geq 0$:

$$\begin{array}{ll} \min_{u,v} & e^T u + e^T v \\ \text{s.t.} & Au - Av = b,\ u \geq 0,\ v \geq 0 \end{array} \tag{3}$$

where e denotes the column vector of 1's and ^{T} denotes transpose. The LP has twice as many variables as (2), but its constraint matrix $\begin{pmatrix} A & -A \end{pmatrix}$ is structured. There has been active recent research showing that, when the columns of A are "nearly orthogonal" (which occurs with high probability when A is randomly generated from, say, a Gaussian distribution) and the soluton of (1) is sufficiently sparse, a solution of (2) also solves (1).

2.2 Robust optimization

The data in an optimization problem may be uncertain, and we would like our solution to be robust in the sense that it is optimal with respect to the worst-case scenario. To illustrate consider an LP:

$$\begin{array}{ll} \min_x & c^T x \\ \text{s.t.} & a_i^T x \leq b_i, \quad i = 1, \ldots, m, \end{array} \tag{4}$$

where $c \in \mathbb{R}^n$, $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$. Suppose a_i is uncertain and all we know is that it belongs to the ellipsoidal set $\mathcal{E}_i = \{\bar{a}_i + A_i u_i \mid \|u_i\|_2 \leq 1\}$, where $\bar{a}_i \in \mathbb{R}^n$, $A_i \in \mathbb{R}^{n \times p_i}$ are given, and $\|\cdot\|_2$ denotes the 2-norm (i.e., $\|u\|_2 = \sqrt{u^T u}$). Then x satisfies $a_i^T x \leq b_i$ under all scenarios $a_i \in \mathcal{E}_i$ means that

$$\max_{a_i \in \mathcal{E}_i} a_i^T x \leq b_i.$$

[2] This can be shown by reduction from the NP-hard integer linear feasibility problem: Given $C \in \mathbb{Z}^{p \times q}$ and $d \in \mathbb{Z}^p$, is there an $x \in \{0,1\}^q$ satisfying $Cx = d$? It can be shown that the answer is 'yes' if and only if the optimal value of (1) equals p, where $A = \begin{pmatrix} A & 0 \\ I & I \end{pmatrix}$, $b = \begin{pmatrix} b \\ e \end{pmatrix}$, I is the identify matrix, and e is the vector of 1's.

We have

$$\max_{a_i \in \mathcal{E}_i} a_i^T x = \max_{\|u_i\|_2 \le 1} (\bar{a}_i + A_i u_i)^T x = \bar{a}_i^T x + \max_{\|u_i\|_2 \le 1} u_i^T (A_i^T x) = \bar{a}_i^T x + \|A_i^T x\|_2,$$

where the last equality uses the observation that a linear function $u \mapsto u^T c$ attains its maximum over the unit Euclidean ball at $u = \frac{c}{\|c\|_2}$, provided $c \neq 0$. Then the robust version of (4) is

$$\begin{array}{ll} \min_x & c^T x \\ \text{s.t.} & \bar{a}_i^T x + \|A_i^T x\|_2 \le b_i, \quad i = 1, \ldots, m. \end{array} \tag{5}$$

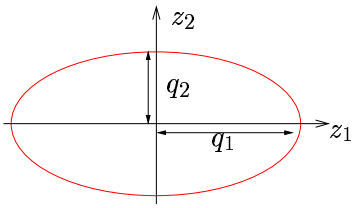
The problem (5) is an example of second-order cone program (SOCP), which is a convex problem and can be solved "efficiently" by, say, interior-point method; also see [1, Section 3.4.2].

2.3 Ellipsoid optimization

For any symmetric $A \in \mathbb{R}^{n \times n}$ ($A = A^T$), we have $\lambda_i(A) \in \mathbb{R}$, $i = 1, \ldots, n$, and

$$\lambda_{\min}(A)\|x\|_2^2 \le x^T A x \le \lambda_{\max}(A)\|x\|_2^2 \quad \forall x \in \mathbb{R}^n.$$

We say A is *positive semidefinite* (abbreviated as "$X \succeq 0$") if $x^T A x \ge 0$ for all $x \in \mathbb{R}^n$ or, equivalently, $\lambda_{\min}(A) \ge 0$. We A is *positive definite* (abbreviated as "$X \succ 0$") if $x^T A x > 0$ for all $0 \neq x \in \mathbb{R}^n$ or, equivalently, $\lambda_{\min}(A) > 0$ or, equivalently, $a_{11} > 0$, $\det \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} > 0$, $\ldots$, $\det A > 0$. For symmetric $A, B \in \mathbb{R}^{n \times n}$, "$A \succeq B$" means $A - B \succeq 0$ and "$A \succ B$" means $A - B \succ 0$.



An ellipse in $\mathbb{R}^2$, centered at the origin and aligned with the coordinate-axes, is specified by the solutions (z_1, z_2) to the following inequality

$$\frac{z_1^2}{q_1} + \frac{z_2^2}{q_2} \le 1,$$

with $q_1 > 0$, $q_2 > 0$. In general, an ellipsoid in $\mathbb{R}^n$, centered at the origin, is specified by the solutions to the inequality

$$z^T Q^{-1} z \le 1,$$

with $Q \in \mathbb{R}^{n \times n}$ being symmetric and positive definite. For n fixed, the volume of the ellipsoid is proportional to $\sqrt{\det Q}$.

In certain applications, we wish to find the minimum volume ellipsoid, centered at the origin, containing a given set of points $z_1, z_2, \ldots, z_p \in \mathbb{R}^n$. This can be formulated as

$$\begin{array}{ll} \min_Q & \det Q \\ \text{s.t.} & Q \succ 0, \quad z_i^T Q^{-1} z_i \le 1, \quad i = 1, \ldots, p. \end{array}$$

This is not friendly looking, however. By taking the natural log of $\det Q$ and making the substitution $X = Q^{-1}$, this can be rewritten as

$$\begin{array}{ll} \min_X & -\log\det X \\ \text{s.t.} & X \succ 0, \quad z_i^T X z_i \leq 1, \quad i = 1, \ldots, p. \end{array} \tag{6}$$

We will see that the problem (6) is a convex problem, which can be solved "efficiently" by, say, interior-point method. There are many variants of this problem, e.g., volume is replaced by the diameter, which is proportional to $\sqrt{\lambda_{\max}(Q)}$. Another is finding the largest ellipsoid, centered at the origin, contained in a polyhedron specified by linear inequalities. See [4, Chapter 5].

2.4 Stability of linear differential equations

Consider the linear ordinary differential equation (ODE)

$$\dot{x}(t) = Ax(t) \quad \forall t \geq 0, \tag{7}$$

where $A \in \mathbb{R}^{p\times p}$, and $x(0) \in \mathbb{R}^p$ is given. It is well-known that it has a unique continous solution $x(\cdot)$ (in fact, $x(t) = e^{At}x(0)$). Moreover, $\lim_{t\to\infty} x(t) = 0$ for every $x(0)$ ("asymptotic stability") if and only if all eigevalues of A have negative real part. It can be shown that this is equivalent to the existence of a matrix $P \in \mathbb{R}^{p\times p}$ satisfying the Lyapunov inequality:

$$A^T P + PA \prec 0, \qquad P \succ 0. \tag{8}$$

($M \prec 0$ means $-M \succ 0$). One direction is easy to see. If (8) holds, then letting

$$V(x) = x^T P x,$$

we have $V(x) \geq 0$ and $V(x) = 0$ if and only if $x = 0$ (since $P \succ 0$). Moreover,

$$\begin{array}{rcl} \frac{d}{dt}V(x(t)) & = & \frac{d}{dt}x(t)^T P x(t) \\ & = & \dot{x}(t)^T P x(t) + x(t)^T P \dot{x}(t) \\ & = & (Ax(t))^T P x(t) + x(t)^T P A x(t) \\ & = & x(t)^T (A^T P + PA) x(t) \\ & \leq & \lambda_{\max}(A^T P + PA)\|x(t)\|_2^2 \\ & < & 0 \quad \text{whenever} \quad x(t) \neq 0. \end{array}$$

From this it's not hard to show that $V(x(t)) \downarrow 0$ and $x(t) \to 0$.

The time-invariant ODE (7) generalizes to the following time-varying ODE:

$$\dot{x}(t) = A(t)x(t), \quad A(t) \in \mathrm{Conv}\{A_1, \ldots, A_m\}, \tag{9}$$

where $A_i \in \mathbb{R}^{p\times p}$, $x(0) \in \mathbb{R}^p$ is given, and

$$\mathrm{Conv}\{A_1, \ldots, A_m\} = \{\alpha_1 A_1 + \cdots \alpha_m A_m \mid \alpha_1 \geq 0, \ldots, \alpha_m \geq 0, \alpha_1 + \cdots + \alpha_m = 1\}.$$

Here (9) is assumed to hold for all $t \geq 0$ except on a set of measure 0. When is this ODE asymptotically stable? There is no simple charactization involving only the eigenvalues of $A_1, \ldots, A_m$ when $m > 1$. However, the Lyapunov inequality (8) generalizes nicely, so that (9) is asymptotically stable (over all $x(0)$ and measurable $A(\cdot)$ satisfying (9)) if and only if there exists a $P \in \mathbb{R}^{p \times p}$ satisfying

$$A_i^T P + P A_i \prec 0, \quad i = 1, \ldots, m, \qquad P \succ 0. \tag{10}$$

Since $M \succ 0$ if and only if $\tau I + M \succeq 0$ for some $\tau < 0$, we can verify the existence of such P by solving the following optimization problem

$$\begin{array}{ll} \min_{P,\tau} & \tau \\ \text{s.t.} & \tau I - A_i^T P - P A_i \succeq 0, \quad i = 1, \ldots, m, \qquad P + \tau I \succeq 0 \end{array} \tag{11}$$

and check if the optimal τ is negative (assuming the minimum in (11) is attained). The problem (11) is an example of a *semidefinite program* (SDP), which is a convex problem and can be solved "efficiently" by, say, interior-point method. Other variants of this problem are discussed in [4, Chapter 5].

2.5 Combinatorial optimization

A well-known intractable problem is the MaxCut problem. In this problem, we are given an undirected graph $\mathcal{G}$ with node set $\mathcal{N} = \{1, \ldots, n\}$, edge set $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ with a nonnegative weight w_{ij} for each edge $(i, j) \in \mathcal{E}$ (so $(i, j) = (j, i)$ and $w_{ij} = w_{ji}$). We wish to find a cut of maximum weight, i.e.,

$$\max_{\mathcal{S} \subseteq \mathcal{N}} w(\mathcal{S}) := \sum_{\substack{(i,j) \in \mathcal{E} \\ i \in \mathcal{S}, j \notin \mathcal{S}}} w_{ij} \ .$$

In the case of $w_{ij} = 1$ for all $(i, j) \in \mathcal{E}$, this seeks a cut with most number of edges. This problem is known to be NP-hard even in this special case. (See, e.g., the 1979 book: Computers and Intractability: A Guide to the Theory of NP-Completeness, by Garey and Johnson.)

A brute-force way to find an optimal subset $\mathcal{S}^*$ is to enumerate all subsets of $\mathcal{N}$, but that would take exponential time as there are 2^n subsets. A 0.5-optimal subset can be found by a simple randomized algorithm: include each node in $\mathcal{S}$ with probability $\frac{1}{2}$, independent of the other nodes. Then each $(i, j) \in \mathcal{E}$ has probability $\frac{1}{2}$ of being in the cut, so the expected weight of the cut is

$$\mathrm{E}[w(\mathcal{S})] = \sum_{(i,j) \in \mathcal{E}} w_{ij} \ \mathrm{P}((i,j) \text{ in the cut}) = \frac{1}{2} \sum_{(i,j) \in \mathcal{E}} w_{ij} \geq \frac{1}{2} \sum_{\substack{(i,j) \in \mathcal{E} \\ i \in \mathcal{S}^*, j \notin \mathcal{S}^*}} w_{ij} = \frac{1}{2} w(\mathcal{S}^*).$$

By repeated trials and saving the cut with the highest weight, we can find (with probability approaching 1) a cut whose weight is within a multiplicative factor of 0.5 of the optimal value. There is also a determinstc greedy algorithm that achieves the same approximation bound of 0.5. For a while, 0.5 was the best bound known – until 1994, when Michel Goemans (MIT) and David Williamson (Cornell) showed using an elegant argument that an SDP relaxation of the MaxCut problem, proposed by Shor and independently by Lovász and Schrijver, yields a bound of 0.87856... Their idea has subsequently been applied to many other NP-hard problems such as

graph partitioning, graph coloring; see [5]. The proof is remarkably simple. We now describe the SDP relaxation and give a proof of this result.

We associate each $\mathcal{S} \subseteq \mathcal{N}$ an $x \in \{-1, 1\}^n$ defined by

$$x_i = \begin{cases} 1 & \text{if } i \in \mathcal{S} \\ -1 & \text{if } i \notin \mathcal{S} \end{cases}$$

Then

$$w(\mathcal{S}) = \sum_{(i,j)\in\mathcal{E}} w_{ij} \left(\frac{1 - x_i x_j}{2} \right) = \sum_{i,j\in\mathcal{N}} c_{ij}(1 - x_i x_j),$$

where we let $c_{ij} = \frac{1}{4} w_{ij}$ for $(i, j) \in \mathcal{E}$ and $c_{ij} = 0$ for $(i, j) \notin \mathcal{E}$. Then the MaxCut problem can be written as as

$$\max_{x\in\{-1,1\}^n} \sum_{i,j\in\mathcal{N}} c_{ij}(1 - x_i x_j). \tag{12}$$

Now, introduce the matrix $X = xx^T$. Notice that $X \succeq 0$ and has rank 1. Moreover, $x_{ij} = x_i x_j$ and $x_i \in \{-1, 1\}$ if and only if $x_{ii} = 1$. Thus, we can rewrite (12) as

$$\begin{array}{ll} \max_X & \sum_{i,j\in\mathcal{N}} c_{ij}(1 - x_{ij}) \\ \text{s.t.} & X \succeq 0, \quad x_{ii} = 1 \ \forall i, \quad \text{rank} X = 1. \end{array}$$

We now *drop the rank-1 constraint*, yielding

$$\begin{array}{ll} \max_X & \sum_{i,j\in\mathcal{N}} c_{ij}(1 - x_{ij}) \\ \text{s.t.} & X \succeq 0, \quad x_{ii} = 1 \ \forall i. \end{array} \tag{13}$$

The problem (13) is a relaxation of MaxCut and is an example of an SDP. We will see that it is a convex problem and can be solved "efficiently" by, say, interior-point method. For now, we will simply assume that we have found an optimal solution X^* of (13), which can be shown to exist due to the compactness of the feasible set of (13) ($X \succeq 0$ implies every principal submatrix of X is positive semidefinite, so $\begin{pmatrix} x_{ii} & x_{ij} \\ x_{ij} & x_{jj} \end{pmatrix} \succeq 0$ for all i, j; since $x_{ii} = x_{jj} = 1$, this implies $|x_{ij}| \leq 1$). Then $X^* \succeq 0$ and $|x^*_{ij}| \leq 1$ for all i, j.

We will now construct from X^* an approximate solution of MaxCut. We will construct this solution randomly, but using a probability distribution based on X^*. Here we follow a Gaussian randomization technique of Bertsimas and Ye instead of the original random-hyperplane approach of Goemans and Williamson. (The Gaussian randomization is simpler to describe and more broadly applicable.) Let $\xi \in \mathbb{R}^n$ be a normal random vector with 0 mean and covariance matrix X^*, i.e., $\xi \sim N(0, X^*)$. Let

$$x_i = \begin{cases} 1 & \text{if } \xi_i > 0 \\ -1 & \text{if } \xi_i \leq 0 \end{cases} \quad \forall i$$

and $S = \{i \mid x_i = 1\}$. Thus

$$\begin{aligned}
\mathrm{E}[w(S)] &= \mathrm{E}\left[\sum_{i,j\in\mathcal{N}} c_{ij}(1 - x_i x_j)\right] \\
&= \sum_{i,j\in\mathcal{N}} c_{ij}\left(1 - \mathrm{E}[x_i x_j]\right) \\
&= \sum_{i,j\in\mathcal{N}} c_{ij}\left(1 - \frac{2}{\pi}\sin^{-1}(x^*_{ij})\right) \\
&\geq \sum_{i,j\in\mathcal{N}} c_{ij}\ (1 - x^*_{ij})\ 0.87856 \\
&\geq w(\mathcal{S}^*)\ 0.87856,
\end{aligned}$$

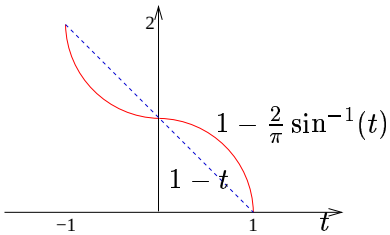
where the third equality uses a result of Sheppard on normal random variables, namely,[3]

$$\mathrm{E}[x_i x_j] = \mathrm{P}[\xi_i \xi_j > 0] - \mathrm{P}[\xi_i \xi_j \leq 0] = \frac{2}{\pi}\sin^{-1}(x^*_{ij}),$$

and the first inequality follows from $c_{ij} \geq 0$, $|x^*_{ij}| \leq 1$, and

$$\min_{-1\leq t\leq 1} \frac{1 - \frac{2}{\pi}\sin^{-1}(t)}{1 - t} =_{\text{by calculus}} 0.87856...$$

(Intuitively, the function $1 - \frac{2}{\pi}\sin^{-1}(t)$ is closely approximated by $1 - t$ for $|t| \leq 1$. We can see this from their graphs shown below.) The last inequality is because (13) is a relaxation of MaxCut (12) (so the optimal value of the former is greater than or equal to that of the latter). By repeated trials and saving the cut with the highest weight, we can find (with probability approaching 1) a cut whose weight is within a multiplicative factor of 0.87856.. of the optimal value. It is not known if the bound of 0.87856.. for (13) can be further improved. It has been shown by Håstad that no polynomial-time algorithm can achieve a bound greater than 0.94117, unless P=NP. Thus, whatever improvement is unlikely to exceed 0.94117. BTW, it is not known if (13) is in the class NP even, although (13) can be solved to high accuracy "efficiently".



• Open question: Can some refinement or variant of the SDP relaxation (13) yield a better bound (i.e., closer to 1)?

[3] see, e.g., page 95 in the 1972 book of Johnson and Kotz: *Distributions in Statistics: Continuous Multivariate Distributions.*