CS292F Convex Optimization: Concepts, Algorithms and Analysis **Spring 2020**

Lecture 2: Optimization basics (April 1)

Lecturer: Yu-Xiang Wang *Scribes: Aiwen Xu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

2.1 Optimization terminology

A *convex optimization problem (or program)* is

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, i = 1, ..., m \\ & Ax = b \end{aligned} \tag{2.1}$$

where f amd $g_i, i = 1, ..., m$ are all convex. The *optimization domain* is $D = \mathrm{dom}(f) \cap_{i=1}^m \mathrm{dom}(g_i)$. f is called *criterion* or *objective* function. g_i is called *inequaltiy constraint* function.

If $x \in D$, $g_i(x) \leq 0, i = 1, ..., m$, and $Ax = b$, then x is called a *feasible point*. The minimum of $f(x)$ over all feasible points x is called *optimal value* f^*. If x is feasible and $f(x) = f^*$, then x is called *optimal*, or a *solution* or *minimizer*. If x is feasible and $f(x) \leq f^* + \epsilon$, then x is called *ϵ-suboptimal*. If x is feasible and $g_i(x) = 0$, g_i is *active* at x.

Lemma 2.1. *Let X_{opt} be the set of all solutions of a convex optimization problem. Then X_{opt} is a convex set. If f is strictly convex, then there exists at most one solution.*

Proof. If $X_{\mathrm{opt}} = \emptyset$, then X_{opt} is trivially convex. If not, let $x_1, x_2 \in X_{\mathrm{opt}}$. Consider $tx_1 + (1-t)x_2, \forall t, 0 \leq t \leq 1$.

$g_i(tx_1 + (1-t)x_2) \leq tg_i(x_1) + (1-t)g_i(x_2) \leq 0$.

$A(tx_1 + (1-t)x_2) = tAx_1 + (1-t)Ax_2 = tb + (1-t)b = b$.

$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) = tf^* + (1-t)f^* = f^*$.

$tx_1 + (1-t)x_2 \in X_{\mathrm{opt}}$ and therefore X_{opt} is a convex set.

Now suppose $X_{\mathrm{opt}} \neq \emptyset$ and f is strictly convex. Suppose there exist $x_1, x_2 \in X_{\mathrm{opt}}$ and $x_1 \neq x_2$. Then by the convexity of X_{opt}, $\frac{1}{2}x_1 + \frac{1}{2}x_2 \in X_{\mathrm{opt}}$. However,

$$f(\frac{1}{2}x_1 + \frac{1}{2}x_2) < \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2) = \frac{1}{2}f^* + \frac{1}{2}f^* = f^*,$$

which is a contradiction to the optimality of f^*. □

Example 2.2. *Lasso.*

Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$,

$$\begin{aligned} \min_{\beta} \quad & \|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \leq s \end{aligned}$$

is a convex optimization problem.

The criterion function $f(\beta) = \|y - X\beta\|_2^2$ is a convex function (least square loss). The inequality constraint is $g(\beta) = \|\beta\|_1 - s$ is a convex function (norm minus a constant). The feasible set is $\{v | \|v\|_1 \leq s, v \in\in \mathbb{R}^p\}$. The solution is unique when $n \geq p$ and X has full column rank, because $X^T X$ is positive definite and therefore f is strictly convex. However, when $p > n$, $X^T X$ is not positive definite. The criterion can be changed to Huber loss

$$\sum_{i=1}^{n} \rho(y_i - x_i^T \beta), \rho(z) = \begin{cases} \frac{1}{2} z^2 & |z| \leq \delta \\ \delta |z| - \frac{1}{2}\delta^2 & |z| > \delta \end{cases},$$

but $\rho(z)$ is not strictly convex. ◁

Example 2.3. *Support vector machine.*

Given $y \in -1, 1^n, X \in \mathbb{R}^{n \times p}$ with rows $x_1, ..., x_n$,

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, i = 1, ..., n \\ & y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, i = 1, ..., n \end{aligned} \tag{2.2}$$

is a convex optimization problem for reasons similar to the above.

The criterion function is not strictly convex in β_0 or ξ. The criterion function is strictly convex in β, so the β component at the solution is unique. If the criterion function is changed to

$$\frac{1}{2}\|\beta\|_2^2 + \frac{1}{2}\beta_0^2 + C \sum_{i=1}^{n} \xi_i^{1.01},$$

then the criterion function is strictly convex in β, β_0 and ξ and the solution is unique. ◁

For a convex problem, a feasible point x is called *locally optimal* when there is some $R > 0$ such that $f(x) \leq f(y)$ for all feasible y such that $\|x - y\|_2 \leq R$.

Proposition 2.4. *For convex optimization problems, local optima are global optima.*

Proof. Suppose that x is a local optimum for some radius $R > 0$. Now suppose x is not a global optimum, i.e., there exists a feasible y such that $f(y) < f(x)$. Then $\|x - y\|_2 > R$. Consider $z = (1 - \theta)x + \theta y, \theta = R/((2\|x - y\|_2))$. By the convexity of the feasible set, z is feasible. And $\|x - z\|_2 = R/2 < R$. Therefore, since x is a local optimum for radius R, it should be the case that $f(x) \leq f(z)$. But because f is convex and $f(y) < f(x)$, $f(z) \leq (1 - \theta)f(x) + \theta f(y) < f(x)$. This is a contradiction. □

We can rewrite the convex optimization problem in (2.1) as follows

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & x \in C = \{x | g_i(x) \leq 0, i = 1, ..., m, Ax = b\}. \end{aligned}$$

Note that C is a convex set.

We can rewrite it further with I_C (the indicator function of C) into an unconstrained form, $\min_x f(x) + I_C(x)$.

2.2 Optimality conditions

The following is called the *first-order condition for optimality.*

Proposition 2.5. *For a convex problem* $\min_x f(x)$ *subject to* $x \in C$*, if* f *is differentiable, a feasible point* x *is optimal if and only if* $\nabla f(x)^T(y-x) \geq 0$ *for all* $y \in C$*.*

Proof. Suppose $x \in C$ and $\nabla f(x)^T(y-x) \geq 0$ for all $y \in C$. Choose any $y \in C$. Since f is convex and differentiable, $f(y) \geq f(x) + \nabla f(x)^T(y-x) \geq f(x)$. So x is optimal.

Suppose x is optimal. Suppose $\nabla f(x)^T(y-x) < 0$ for some $y \in C$. Consider $z(t) = (1-t)x + ty, t \in [0,1]$. $z(t) \in C$ due to the convexity of C. Since

$$\lim_{t \to 0} \frac{f(z(t)) - f(x)}{t} = \nabla f(x)^T(y-x) < 0.$$

Therefore, for some small positive t, $f(z(t)) < f(x)$, which contradicts the optimality of x. □

If $C = \mathbb{R}^n$, then the condition reduces to $\nabla f(x) = 0$. $C = \mathbb{R}^n$ implies that $\nabla f(x)^T v \geq 0$ holds for all $v \in \mathbb{R}$. For a particular nonzero v, $\nabla f(x)^T v \geq 0$ and $\nabla f(x)^T(-v) \geq 0$. So $\nabla f(x) = 0$.

Example 2.6. *Quadratic minimization.*

Consider minimizing

$$f(x) = \frac{1}{2}x^TQx + b^Tx + c, Q \succeq 0$$

over $\mathbb{R}^n$. The first-order optimality condition says that the solution satisfies

$$\nabla f(x) = Qx + b = 0.$$

If $Q \succ 0$, then there is a unique solution $x = -Q^{-1}b$.

If Q is singular and $b \notin \mathrm{col}(Q)$, then there is no solution, i.e., $\min_x f(x) = -\infty$.

If Q is singular and $b \in \mathrm{col}(Q)$, then there are infinitely many solutions $x = -Q^+b + z, z \in \mathrm{null}(Q)$, where Q^+ is the pseudoinverse of Q. ◁

Example 2.7. *Equality-constrained minimization.*

Consider

$$\min_x f(x) \text{ subject to } Ax = b,$$

where f is differentiable. According to the first-order optimality condition, x is optimal if $Ax = b$ and

$$\nabla f(x)^T(y-x) \geq 0, \forall y : Ay = b.$$

For any y, $y = x + v$ for some $v \in \mathrm{null}(A)$. Therefore the condition becomes $\nabla f(x)^T v \geq 0, \forall v \in \mathrm{null}(A)$. If a linear function is nonnegative on a subspace, then it must be zero on the subspace. Then $\nabla f(x)^T v = 0, \forall v \in \mathrm{null}(A)$, i.e., $\nabla f(x) \perp \mathrm{null}(A)$. Since $\mathrm{null}(A)^\perp = \mathrm{row}(A^T)$, the condition becomes $\nabla f(x) \in \mathrm{row}(A^T)$, i.e.,

$$\exists v, \nabla f(x) + A^T v = 0,$$

which is known as *Lagrange multiplier optimality condition.* ◁

Example 2.8. *Projection onto a convex set.*

Given a and a convex set C, consider

$$\min_x \|a-x\|_2^2 \text{ subject to } x \in C.$$

By the first-order optimality condition, x satisfies

$$\nabla f(x)^T(y-x) = (x-a)^T(y-x) \geq 0, \forall y \in C.$$

Rewrite this equation into $(a-x)^T x \geq (a-x)^T y$, and recall that the *normal cone* to C at x is $\mathcal{N}_C(x) = \{g : g^T x \geq g^T y, \forall y \in C\}$. So the condition becomes $a - x \in \mathcal{N}_C(x)$. ◁

2.3 Equivalent transformations

Informally, we call two problems *equivalent* if from a solution of one, a solution of the other is readily found, and vice versa. The following transformation techniques convert an optimization problem into an equivalent optimization problem.

2.3.1 Partial optimization

Partial optimization preserves the convexity of a problem. Consider the convex optimization problem

$$\begin{array}{ll} \min\limits_{x_1,x_2} & f(x_1,x_2) \\ \text{subject to} & g_1(x_1) \leq 0, g_2(x_2) \leq 0. \end{array}$$

If we decompose $x = (x_1, x_2) \in \mathbb{R}^{n_1+n_2}$, then the problem

$$\begin{array}{ll} \min\limits_{x_1} & \tilde{f}(x_1,x_2) = \min\{f(x_1,x_2) : g_2(x_2) \leq 0\} \\ \text{subject to} & g_1(x_1) \leq 0. \end{array}$$

is also convex.

Example 2.9. *Hinge form of SVM*

Rewrite the constraints in the SVM problem (2.2) as

$$\xi_i \geq \max\{0, 1 - y_i(x_i^T\beta + \beta_0)\}.$$

At the optimal solution, we must have those constraints active, i.e., the equality must be achieved. (Suppose the constraints are not active at the optimal solution, i.e., we have $\xi_i > \max\{0, 1 - y_i(x_i^T\beta + \beta_0)\}$. We can decrease ξ_i without violating the constraints and thus decrease the objective, which is a contradiction to the optimality of the solution.) If we plug in the values of the optimal ξ_i, we have the *hinge form* of SVM,

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n [1 - y_i(x_i^T\beta + \beta_0)]_+,$$

where $a_+ = \max\{0, a\}$ is called the *hinge function*. ◁

2.3.2 Transformation and change of variables

Let $h : \mathbb{R} \to \mathbb{R}$ be a *monotonically increasing transformation.* Then

$$\min_x f(x) \text{ subject to } x \in C \iff \min_x h(f(x)) \text{ subject to } x \in C.$$

Similarly, inequality or equality constraints can be transformed and yield equivalent optimization problems. If $phi : \mathbb{R}^n \to \mathbb{R}^m$ is one-to-one, and its image covers the feasible set C, then we can $change variables$, i.e.,

$$\min_x f(x) \text{ subject to } x \in C \iff \min_y f(\phi(y)) \text{ subject to } \phi(y) \in C.$$

Example 2.10. *Geometric programming.*

A *monomial* is a function $f : \mathbb{R}^n_{++} \to \mathbb{R}$ of the form $f(x) = \gamma x_1^{a_1} x_2^{a^2} \cdots x_n^{a_n}$ where $\gamma > 0, a_1, ..., a_n \in \mathbb{R}$. A *posynomial* is a sum of monomials, $f(x) = \sum_{k=1}^p \gamma_k x_1^{a_{k1}} x_2^{a_{k2}} \cdots x_n^{a_{kn}}$. A *geometric program* is of the form

$$\begin{array}{ll} \min\limits_x & f(x) \\ \text{subject to} & g_i(x) \leq 1, i = 1, ..., m \\ & h_j(x) = 1, j = 1, ..., r \end{array}$$

where $f, g_i, i = 1, ..., m$ are posynomials and $h_j, j = 1, ..., r$ are monomials. This is nonconvex.

Given $f(x) = \gamma x_1^{a_1} x_2^{a^2} \cdots x_n^{a_n}$, let $y_i = \log x_i$ and rewrite this as $\gamma(e^{y_1})^{a_1}(e^{y_2})^{a_2} \cdots (e^{y_n})^{a_n} = e^{a^T y + b}$ where $b = \log \gamma$. A posynomial can be rewritten as $\sum_{k=1}^p e^{a_k^T y + b_k}$. After taking logarithms, a geometric program is equivalent to

$$\begin{array}{ll} \min\limits_x & \log(\sum\limits_{k=1}^{p_0} e^{a_{0k}^T y + b_{0k}}) \\ \text{subject to} & \log(\sum\limits_{k=1}^{p_i} e^{a_{ik}^T y + b_{ik}}) \leq 0, i = 1, ..., m \\ & c_j^T y + d_j = 0, j = 1, ..., r, \end{array}$$

which is a convex optimization problem. ◁

2.3.3 Eliminating equality constraints

Given a convex optimization problem (2.1), any feasible point can be expressed as $x = My + x_0$, where $Ax_0 = b$ and $\mathrm{col}(M) = \mathrm{null}(A)$. Equivalently, the problem becomes

$$\begin{array}{ll} \min\limits_y & f(My + x_0) \\ \text{subject to} & g_i(My + x_0) \leq 0, i = 1, ..., m. \end{array}$$

2.3.4 Introducing slack variables

Given a convex optimization problem (2.1), we can transform the inequality constraints as follows

$$\begin{array}{ll} \min_{x,s} & f(x) \\ \text{subject to} & s_i \geq 0, i = 1, ..., m \\ & g_i(x) + s_i = 0, i = 1, ..., m \\ & Ax = b. \end{array}$$

This problem is no longer convex unless $g_i, i = 1, ..., n$ are affine.

2.3.5 Relaxation

Given an optimization problem

$$\min_x f(x) \text{ subject to } x \in C,$$

we can always take an enlarged constraint set $\tilde{C} \supseteq C$ and consider

$$\min_x f(x) \text{ subject to } x \in \tilde{C}.$$

This is called a *relaxation* and its optimal value is always smaller or equal to that of the original problem.

An important special case is *relaxing nonaffine equality constraints*, i.e.,

$$h_j(x) = 0, j = 1, ..., r,$$

where $h_j, j = 1, ..., r$ are convex but nonaffine, are replaced with

$$h_j(x) \leq 0, j = 1, ..., r.$$

Example 2.11. *Maximum utility problem.*

The *maximum utility problem* models investment or consumption

$$\begin{array}{ll} \max_{x,b} & \sum_{t=1}^{T} \alpha_t u(x_t) \\ \text{subject to} & b_{t+1} = b_t + f(b_t) - x_t, t = 1, ..., T \\ & 0 \leq x_t \leq b_t, t = 1, ..., T, \end{array}$$

where b_t is the budget, x_t is the amount consumed at time t, f is an investment return function, u is an utility function, and f and u are concave and increasing.

This is not a convex optimization problem because the equality constraints are not affine in b. If we relax the equality constraints to

$$b_{t+1} \leq b_t + f(b_t) - x_t, t = 1, ..., T,$$

which becomes a convex optimization problem. ◁

Example 2.12. *Principal component analysis.*

Given $X \in \mathbb{R}^{n \times p}$, consider

$$\min_R \|X - R\|_F^2 \text{ subject to } \operatorname{rank}(R) = k.$$

This is not convex because of rank.

Given $X = U\Sigma V^T$ (the singular value decomposition of X), the solution is $R = U\Sigma_k V^T$, where Σ_k is a diagonal matrix with the first k diagonal elements of Σ. R is said to be a reconstruction of X from its *first k principal components.*

Now we rewrite this problem into the following equivalent form (explained in [1])

$$\min_Z \|X - XZ\|_F^2 \text{ subject to } \operatorname{rank}(Z) = k, Z^2 = Z, Z = Z^T.$$

The solution to this problem is $Z = V_k V_k^T$ where V^k contains the first k columns of V and the rest of the entries are 0.

Now the feasible set can be written as

$$C = \{Z \in \mathbb{S}^p : \lambda_i(Z) \in \{0, 1\}, i = 1, ..., p, \operatorname{tr}(Z) = k\},$$

where $\lambda_i(Z)$ are the eigenvalues of Z and $\mathbb{S}^p$ is the set of symmetric $p \times p$ matrices. We can now relax this set into its convex hull

$$\begin{aligned} \mathcal{F}_k &= \{Z \in \mathbb{S}^p : \lambda_i(Z) \in [0, 1], i = 1, ..., p, \operatorname{tr}(Z) = k\} \\ &= \{Z \in \mathbb{S}^p : 0 \succeq Z \succeq I, \operatorname{tr}(Z) = k\}. \end{aligned}$$

$\mathcal{F}_k$ is called the *Fantope* of order k, and it's a convex set. Now the problem becomes

$$\min_Z \|X - XZ\|_F^2 \text{ subject to } Z \in \mathcal{F}_k.$$

Note that

$$\|X - XZ\|_F^2 = \operatorname{tr}((X - XZ)^T(X - XZ)) = \operatorname{tr}(X^T X - Z^T X^T X) = \operatorname{tr} X^T X - \operatorname{tr}(X^T XZ),$$

so the minimization of $\|X - XZ\|_F^2$ is equivalent to the maximization of $\operatorname{tr}(X^T XZ)$. Now the problem becomes

$$\max_Z \operatorname{tr}(X^T XZ) \text{ subject to } Z \in \mathcal{F}_k,$$

which is a convex problem because it is the maximization of a linear objective over a convex set.

The advantage of an optimization formulation is that the problem can be modified to address additional problem specific considerations. If we want the recovered principal components to be sparse, we can modified the problem into

$$\max_Z \operatorname{tr}(X^T XZ) - \lambda \sum_{i,j} |Z_{i.j}| \text{ subject to } Z \in \mathcal{F}_k,$$

which is the problem considered in [2]. ◁

Example 2.13. *Approximate algorithm for Max-Cut.*

Given a graph with nodes, edges and edge weights, we want to find a subset S of the nodes such that the sum of the weights w_{ij} of the edges between S and its complement $\bar{S}$ is maximized.

We can formulate the problem as follows. Let $x_j = 1$ if $j \in S$ and $x_j = -1$ if $j \in \bar{S}$. Then

$$\begin{aligned} \max_x \quad & \frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n w_{ij}(1 - x_i x_j) \\ \text{subject to} \quad & x_j \in \{-1, 1\}, j = 1, ..., n. \end{aligned}$$

Let $Y = xx^T$. Note that $Y_{i,i} = 1, i = 1, ..., n$. Rewriting the above problem, we have

$$\begin{array}{ll} \max\limits_{x} & \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-(xx^T)_{i,j}) \\ \text{subject to} & x \in \{-1,1\}^n. \end{array}$$

Relaxing the problem into a convex problem, we have

$$\begin{array}{ll} \max\limits_{Y\in\mathbb{R}^{n\times n}} & \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-Y_{i,j}) \\ \text{subject to} & Y_{i,i} = 1, i = 1, ..., n \\ & Y \succeq 0. \end{array}$$

Then we sample v uniformly from the unit sphere in $\mathbb{R}^n$ and output $\text{sign}(Yv)$. This can be shown to be a 0.87856 approximation of this NP-complete problem. It is known as *Geomans and Williamson algorithm.* $\triangleleft$

References

[1] Madeleine Udell, "Generalized Low Rank Models," *Stanford University Thesis*, `https://people.orie.cornell.edu/mru8/doc/udell15_thesis.pdf`.

[2] Vincent Q. Vu, Juhee Cho, Jing Lei, Karl Rohe, "Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA," *Advances in Neural Information Processing Systems 26 (NIPS 2013).*