Optimization

David Connelly
Spring 2022

Contents

1 Convexity

1.1 Convex sets and functions

A set S is convex if, for any x and y in S and any $\theta \in [0,1]$, the convex combination $\theta x + (1-\theta y)$ is also in S. The set S is a cone if for any $x \in S$ and $\theta \geq 0$ we have $\theta x \in S$ as well. Two important convex cones are

- the second order cone $\big\{(x,t) \in \mathbb{R}^{n+1} \mid \|x\|_2 \leq t\big\}$
- the cone of positive semi-definite matrices $\big\{X \in \mathbb{R}^{n\times n} \mid X = X^\top \text{ and } X \geq 0\big\}$

Convexity is preserved under intersection and under application of an affine function — that is, a function of the form $f(x) = Ax + b$.

A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex if its domain is convex and for any x and y in the domain and any $\theta \in [0,1]$ we have

$$f\Big(\theta x + (1-\theta)y\Big) \leq \theta f(x) + (1-\theta)f(y)$$

Note that the latter condition is equivalent to the epigraph of f being a convex set. For the remainder of this discussion we will not explicitly state the condition that f has a convex domain. The function f is concave if $-f$ is convex.

If f is differentiable, we can characterize its convexity using information about its derivatives. First, f is convex if and only if for any x and y in the domain we have

$$f(x) + \nabla f(x)^\top (y-x) \leq f(y)$$

Alternatively, f is convex if and only if

$$\nabla^2 f(x) \geq 0$$

for all x in the domain, where $\nabla^2 f(x)$ is the Hessian matrix.

1.2 Separating hyperplanes

A hyperplane is a set of the form $\big\{x \in \mathbb{R}^n \mid a^\top x = b\big\}$ for fixed vector a and scalar b. Hyperplanes divide $\mathbb{R}^n$ into two half-spaces. The separating hyperplane theorem states that if C and D are convex sets with $C \cap D = \emptyset$, then there exist a vector a and scalar b such that $a^\top x \leq b$ for all $x \in C$ and $a^\top x \geq b$ for all $x \in D$.

The proof is worth knowing mainly in the special case where the distance between C and D is positive and attained by points $c \in C$ and $d \in D$. Below is a sketch of the proof under those mild assumptions.

1. Define the parameters of the hyperplane

$$a = d - c \text{ and } b = \frac{d^\top d - c^\top c}{2}$$

2. Assume there exists a $u \in D$ with $a^\top u < b$ and show this implies $(d-c)^\top (u-d) < 0$.
3. Show that the derivative of $\|d + t(u-d) - c\|^2$ with respect to t at $t = 0$ is negative, such that for small $t > 0$ we have $d + t(u-d)$ closer to c than d was, contradicting the choice of c and d.

If a and b can be found such that the inequalities in the theorem are strict, we say that the sets are strictly separated.

1.3 Strong convexity

In this section we will fix a point x_0 and restrict f to the sublevel set $S = \{x \in \operatorname{dom} f \mid f(x) \leq f(x_0)\}$. We will assume that S is closed.

The function f is strongly convex on S if there exists an $m > 0$ such that $\nabla^2 f(x) \geq mI$ for all $x \in S$. Strong convexity provides a few useful inequalities. First, for all x and y in S we have

$$f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2} \|y - x\|^2 \leq f(y)$$

which can be shown through Taylor expansion.

Next, note that the previous inequality means that $\|x - y\|$ must be bounded on S, because $f(y) \leq f(x_0)$ and so we cannot be able to make the left-hand side arbitrarily large. Thus S is bounded. Moreover, the function that maps $\nabla^2 f(x)$ to its largest eigenvalue is continuous in x, so it must achieve a maximum on the closed and bounded set S. It follows that there is an $M > 0$ such that $\nabla^2 f(x) \leq MI$ for all $x \in S$. The ratio $\kappa = M/m$ bounds the condition number of the Hessian. We have the corresponding inequality

$$f(y) \leq \nabla f(x)^\top (y - x) + \frac{M}{2} \|y - x\|^2$$

These inequalities can be leveraged to bound $\|x - x^*\|$ and $f(x) - f(x^*)$ in a way that is rather straightforward.

1.4 Subgradients and subdifferentials

A vector y is a subgradient to a convex function f at x if $f(x) + y^\top z \leq f(x + z)$ for all vectors z. The set of all such subgradients is known as the subdifferential and may be denoted $\partial f(x)$.

Recall that the conjugate function is

$$f^*(y) = \sup_z \left\{ y^\top z - f(z) \right\}$$

Then the Fenchel-Young theorem states that

$$x^\top y \leq f(x) + f^*(y)$$

with equality if and only if $y \in \partial f(x)$.

We also define the directional derivative

$$f'(x, d) = \lim_{t \to 0^+} \frac{f(x + td) - f(x)}{t}$$

1.5 Standard problems

The general form of the convex optimization problems we consider is

$$\begin{aligned} \text{minimize } & f_0(x) \\ \text{subject to } & f_i(x) \leq 0, \ 1 \leq i \leq m \\ & a_i^\top x = b_i, \ 1 \leq i \leq p \end{aligned}$$

where f_i must be convex for $0 \leq i \leq m$. If f_0 is differentiable, then a necessary and sufficient condition for a feasible x to be optimal is

$$\nabla f_0(x)^\top (y - x) \geq 0$$

for all feasible y. If there are no constraints, this reduces to $\nabla f_0(x) = 0$.

One specific convex problem is the linear program in standard form, given below.

$$\begin{aligned} \text{minimize } & c^\top x \\ \text{subject to } & Ax = b \\ & x \geq 0 \end{aligned}$$

Another common form is the semi-definite program

$$\begin{aligned} \text{minimize} \ & \operatorname{tr}(CX) \\ \text{subject to} \ & \operatorname{tr}(A_iX) = b_i, \ 1 \leq i \leq p \\ & X \geq 0 \end{aligned}$$

where X is required to be symmetric.

2 Duality

2.1 The Lagrangian dual function

We consider a (not necessarily convex) optimization problem (so the equality constraints are just $h_i(x) = 0$ for $1 \leq i \leq p$). The Lagrangian $L\colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is given by

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

Then the Lagrangian dual function $g\colon \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$$

where $\mathcal{D}$ is the feasible domain.

Denote the optimal value of the objective function f_0 by p^*, and let the optimal value be achieved at x^*. Take any $\lambda \geq 0$ and any ν. Then for any feasible x we have

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \leq f_0(x)$$

so that

$$\begin{aligned} g(\lambda, \nu) &= \inf_{y \in \mathcal{D}} L(y, \lambda, \nu) \\ &\leq L(x, \lambda, \nu) \\ &\leq f_0(x) \end{aligned}$$

In particular, the above holds for $x = x^*$, so that $\boxed{g(\lambda, \nu) \leq p^*}$ for any λ and ν chosen as above.

For reference, we write the Lagrangian dual function of the standard linear program explicitly. It is

$$g(\lambda, \nu) = \begin{cases} -b^\top \nu & A^\top \nu - \lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

2.2 The Lagrangian dual problem

The Lagrangian dual of a standard optimization problem is simply

$$\begin{aligned} &\text{maximize } g(\lambda, \nu) \\ &\text{subject to } \lambda \geq 0 \end{aligned}$$

The optimal value d^* of the Lagrangian dual problem satisfies $d^* \leq p^*$, regardless of the convexity of the problem.

2.3 Strong duality and Slater's condition

If $d^* = p^*$ then strong duality holds. For convex minimization problems, one condition that guarantees strong duality is that of Slater. Namely, if there is an $x \in \operatorname{relint} \mathcal{D}$ with $Ax = b$ and $f_i(x) < 0$ for $1 \leq i \leq m$, then strong duality holds. Here $\operatorname{relint} \mathcal{D}$ is defined similarly to the usual interior, but only the portion of each neighborhood intersecting the affine hull needs to be in the set.

3 Convex methods

3.1 Descent methods

Here we consider unconstrained minimization of a strongly convex function f using methods that produce a minimizing sequence

$$x_{k+1} = x_k + t_k \Delta x_k$$

where $f(x_{k+1}) < f(x_k)$. At each iteration, such methods must first determine a search direction Δx_k and subsequently choose a scale factor t_k. Ideally, the latter task would be accomplished by choosing

$$t = \operatorname*{argmin}_{s>0} f(x_k + s\Delta x_k)$$

However, the above problem is often itself expensive to solve. Instead, we can use backtracking line search, which starts by setting $t_k = 1$ and then sets $t = \beta t$ until $f(x_k + t\Delta x_k) \leq f(x_k) + \alpha t_k \nabla f(x_k)^\top \Delta x_k$. Here the constants are freely chosen to satisfy $0 < \alpha < 0.5$ and $0 < \beta < 1$.

3.2 Gradient descent

The gradient descent method consists simply of choosing $\Delta x_k = -\nabla f(x_k)$. Convergence analysis of gradient descent, regardless of the line search method, begins by noting that strong duality gives

$$f\Big(x - t\nabla f(x)\Big) \leq f(x) - t\|\nabla f(x)\|^2 + \frac{Mt^2}{2}\|f(x)\|^2$$

If exact line search is used, we proceed by minimizing both sides of the inequality, and eventually find

$$f(x_k) - p^* \leq \left(1 - \frac{m}{M}\right)^k \big(f(x_0) - p^*\big)$$

which proves convergence and establishes some bounds on the rate. If backtracking line search is used, the analysis is more complicated, so I omit the proof. The result is

$$f(x_k) - p^* \leq c^k \big(f(x_0) - p^*\big)$$
$$\text{where } c = 1 - \min\left\{2\alpha m, \frac{2\alpha\beta m}{M}\right\}$$

3.3 Newton's method

The second-degree Taylor series of f about $\mathbf{x}$

$$f(x + y) \approx f(x) + \nabla f(x)^\top y + \frac{1}{2} y^\top \nabla^2 f(x) y$$

is easily minimized by setting the gradient equal to zero, such that

$$y = -\nabla^2 f(x)^{-1} \nabla f(x)$$

is the minimizer. Newton's method is the gradient method with the above value as the search direction, possibly with backtracking line search to set the step magnitude. Convergence can be shown to occur in under $6 + C$ iterations, where C is some reasonable constant dependent on the smoothness of f.

3.4 Nesterov complexity

Suppose we have a method that updates x_k based on values of f and ∇f, and moreover that increment $x_k - x_0$ is always in the span of the gradients computed thus far. Then Nesterov proved that, for any $0 < m < M$, there exists an f having m and M as strong convexity constants such that

$$\left\|x^k - x^*\right\|^2 \geq \left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\right)^{2k} \left\|x^0 - x^*\right\|^2$$

where x^* is the optimizer and $\kappa = M/m$. This result bounds how good this natural class of methods can be. Note that we have switched to using superscripts for iteration numbers.

We assume we are working in the infinite-dimensional space ℓ^2. Given $m < M$, we define

$$f(x) = \frac{M - m}{8}\left(x_1^2 + \sum_{i=1}^{\infty}(x_i - x_{i+1})^2 - 2x_1\right) + \frac{m}{2}\|x\|^2$$

Define the tridiagonal matrix

$$T = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \end{pmatrix}$$

Then it is straightforwad to compute

$$\begin{aligned} \nabla f(x) &= \left(\frac{M - m}{4}T + mI\right)x - \frac{M - m}{4}e_1 \\ \nabla^2 f(x) &= \frac{M - m}{4}T + mI \end{aligned}$$

In particular, the latter equality shows that f is strongly convex with the constants m and M. To find the minimizer, we set $\nabla f = 0$, which leads to the difference equation

$$x_{i+1} - 2\frac{M + m}{M - m}x_i + x_{i-1} = 0$$

which we solve by guessing $x_i = q^i$ and finding

$$q = \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}$$

From here, we demonstrate that f is the function claimed to exist by Nesterov by taking $x^0 = 0$ without loss of generality and showing that the tridiagonal structure of T means that the iterates only gain components one at a time, so their ℓ^2 norms can be bounded.

3.5 Nesterov's accelerated gradient method

If f is strongly convex, then setting $x^0 = y^0$ and iterating

$$\begin{aligned} x^{k+1} &= y^k - \frac{1}{M}\nabla f\left(y^k\right) \\ y^{k+1} &= x^{k+1} + q\left(x^{k+1} - x^k\right) \end{aligned}$$

achieves Nesterov's lower bound shown above.