CONVEX OPTIMIZATION

MINSEOK SONG

Some facts in Convex optimization

- We have recurring theme, specifically certain inequalities, arising in convex optimization.

Fact 1. *Let S be an open convex set. A function $f : S \to \mathbb{R}$ is convex iff for every $w \in S$, there exists v such that*

$$\forall u \in S, f(u) \geqslant f(w) + \langle u - w, v \rangle$$

Definition 1. *A function f is strongly convex on S, if $\exists m > 0$ such that $\nabla^2 f \succeq mI$.*

In case f is not differentiable, more general definition gives, $\forall w, u$, and $\forall \alpha \in (0, 1)$

$$f(\alpha w + (1-\alpha)u) \leqslant \alpha f(w) + (1-\alpha)f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2.$$

- This is a function that grows as fast as quadratic function.
- We can view $\|w - u\|$ as a "penalty" of the distance between w and u.

Fact 2. *If f is λ-strongly convex then*

$$\forall w, u, \exists v \in \partial f(w), \langle w - u, v \rangle \geqslant f(w) - f(u) + \frac{\lambda}{2}\|w-u\|^2$$

Proof. $f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(z)(y-x)$ for some z on the segment $[x, y]$. By the definition of strong convexity, we have $\nabla^2 f(z) \succeq mI$. □

Gradient Descent, Stochastic Gradient Descent

Fact 3. *If f is λ-strongly convex then $\forall w, u$, and $v \in \partial f(w)$, we have*

$$\langle w - u, v \rangle \geqslant f(w) - f(u) + \frac{\lambda}{2}\|w-u\|^2$$

Proof. Use the definition of strong convexity, along with subgradient. □

- For this type of inequality in general, we use 1) the definition of convexity/strong convexity & limiting argument as $\alpha \to 0$ 2) mean value theorem

Fact 4. *Let v_i's $i = 1, \ldots, T$ be arbitrary sequence of vectors (think of it as the direction of update). Any algorithm with $w^{(1)} = 0$ and the update rule of the form $w^{(t+1)} = w^{(t)} - \eta v_t$ satisfies*

$$\sum_{t=1}^{T}\langle w^{(t)} - w^*, v_t\rangle \leqslant \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2.$$

(1) GD: appropriate for when the function is differentiable.
(2) SGD: even if the function is not differentiable, we can use subgradient. With certain condition, convergence holds.
(3) Two-step SGD: resolve the concern that the norm of w might increase by using projection.
(4) As a generalization, we could make η depend on t, or in error analysis instead of using the average of w^t, we can use the last few elements (other variants exist as well).

(5) Strong convexity: might achieve faster convergence.
(6) β-smooth: when we use loss function, instead of using Lipschitz function, we can impose the condition of β-smoothness and have different convergence.
   (a) Assume that $(\cdot, z)l$ is convex, β-smooth, and nonnegative. Then
$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\bar{w})] \leqslant \frac{1}{1-\eta\beta}(\mathcal{L}_{\mathcal{D}}(w^*) + \frac{\|w^*\|^2}{2\eta T}).$$
   (b) Note that we have a specific choice of $\eta = \frac{1}{\beta(1+3/\epsilon)}$ in order to achieve ϵ error (specifically, $T \geqslant 12B^2\beta/\epsilon^2$).
   (c) The point of β-smooth property of the function is the self-boundedness, i.e. $\|\nabla f(w)\|^2 \leqslant 2\beta f(w)$, with the additional condition that l is nonnegative and convex : here, we prove it by first observing that
$$f(v) \leqslant f(w) + \nabla f(w)^T(v-w) + \frac{\beta}{2}\|v-w\|^2.$$

- Some confusing notation and points:
  (1) $f_t(\cdot) = l(\cdot, w^{(t)})$.
  (2) $f(w) = \mathcal{L}_{\mathcal{D}}(w) = \mathbb{E}_{z\sim\mathcal{D}}[l(w,z)]$. l is a loss function and $\mathcal{L}$ is an expected loss.
  (3) $f_t(w)$ is still a random variable since w is random.

KKT theorem

Theorem 1. *Assume that functions $f_0, \ldots, f_m, h_1, \ldots, h_p$ are differentiable. We are trying to solve the optimization problem*

$$\begin{aligned} &\textit{Minimize } f_0(x) \\ &\textit{subject to } f_i(x) \leqslant 0, \quad i = 1, 2, \ldots, m, \\ &\qquad\qquad h_j(x) = 0, \quad j = 1, 2, \ldots, p. \end{aligned}$$

x^* is a solution of this problem and the strong duality holds if and only if x^* satisfies

$$\begin{aligned} &f_i(x^*) \leqslant 0, \quad i = 1, \ldots, m \\ &h_i(x^*) = 0, \quad i = 1, \ldots, p \\ &\lambda_i^*(x) \geqslant 0, i = 1, \ldots, m \\ &\lambda_i^* f_i(x^*) = 0, i = 1, \ldots, m \\ &\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*) = 0 \\ &\textit{for some } \lambda^* \geq 0 \end{aligned}$$

Proof. ($\Leftarrow$) The key here is to notice that $\lambda \geqslant 0$ because this makes L, a duality function, convex. It follows that x^* gives a minimum solution of L by the last condition. Using the fourth and the second condition, we can see that $f_0(x^*) = g(\lambda^*, \nu^*)$

($\Rightarrow$) This holds by strong duality. □

- $\lambda^T f_i(x) = 0$ is called complemtary slackness. The term complementary comes from the idea of either or that, and slackness means "leftover." This condition is hard to verify in practice.
- Caveat: if the strong duality condition does not hold, then KKT condition might not hold. Further, of course, by no means do we have existence and uniqueness.

- Some of variants exist, one of which includes when f is not convex, given below.

Lemma 2. *Consider now the equality constrained problem, adding the condition*

$$h_i(x) = 0, i = 1, 2, \ldots, p$$

from the previous setup. Assume all functions are continuously differentiable. Let x^* be the global minimum of a problem. Assume that the gradients of $h_i(x), i = 1, 2, \ldots, p$ are linearly independent at x^*. There exist $\nu_1, \nu_2, \ldots, \nu_p \in \mathbb{R}$ (the Lagrange multipliers) such that $\nabla f(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$.

Remark 1.
Note that complementary slackness and linearity of h_i are compensated by independence, with a simpler setup here - in a way, independence is a pretty strong condition here.

Proof. It follows by observing that $\nabla f(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$. By independence, we also have a uniqueness. □

Theorem 3. *Assume further that we have continuously differentiable functions $g_i(x) \leqslant 0, i = 1, 2, \ldots, m$.*
Let us assume that the vector set composed of $\nabla h_i(x)$ and the gradients of all active inequality constraints form a linear independent set. At x^*, the KKT condition holds.

Remark 2.
Gradient is the steepest direction, and independence signifies the well-behavedness of the constraint, in the sense that whenever x^* is in the 'edge' of the inequality constraint, this direction should give a new information. The theorem says that this well-behavedness is sufficient condition for KKT.

Proof.
(1) It is enough to show for the case g_i's are all active at x^*, because $\mu_i = 0$ whenever $g_i \neq 0$ since $\mu_i \geqslant 0$ (otherwise we do not have minimum at x^*).
(2) Since $g_i \leqslant 0$, we can also assume that inequality is equality, and use the previous theorem to show the existence of Lagrangian constants that satisfy $\nabla L = \nabla_x f(x^*) + \sum_{i=1}^{m} \mu_i \nabla_x g_i(x^*) + \sum_{i=1}^{p} \nu_i \nabla_x h_i(x^*) = 0$.
(3) Now it suffices to prove $\mu_i \geqslant 0$ for all i's. Assume, to the contrary, that $\mu_i < 0$ for some i. First note that $\sum_{i=1}^{m} \mu_i \nabla_x g_i(x^*) + \sum_{i=1}^{p} \nu_i \nabla_x h_i(x^*) = -\nabla_x f(x^*) = 0$ since x^* achieves minimum. Use constant rank theorem on $F(t, x) = (g_1(x), \ldots t + g_i(x), \ldots, g_m(x), h_1(x) \ldots, h_p(x))$ (we may need additional condition, such as independence of gradients). We can derive the contradiction by having $\nabla_x f(x^*) = -\sum_{i=1}^{m} \mu_i \nabla_x g_i(x^*) - \sum_{i=1}^{p} \nu_i \nabla_x h_i(x^*) \neq 0$.

□

Remark 3.
Dual is not necessarily symmetric in mathematics in general. For example the dual of L^1 is L^∞ but the dual of L^∞ is not L^1.
Under certain condition, it is; for example if f is convex and closed, (epigraph is a closed set) then $f^{**} = f$.
Another relevant concept for convex analysis (in particular, for function) is Legendre transform.