Math 173B - Lecture 10: Constrained Optimization

1 Optimization under Constraints

1.1 Single equality constraint

Example: Consider the following problem:

$$\begin{aligned} &\text{minimize } x+y, \\ &\text{subject to } x^2+y^2=1. \end{aligned}$$

Let us recall a classical analytical method: Lagrange multipliers.

Define the Lagrangian:

$$L(x,y,\lambda)=x+y+\lambda(x^2+y^2-1).$$

$L(x,y,\lambda)$ has stationary points given by $\nabla L=0$. Here the gradient operator ∇ is taken with respect to all variables x,y,λ. We have

$$\nabla L(x,y,\lambda)=\begin{pmatrix}\partial_x L\\ \partial_y L\\ \partial_\lambda L\end{pmatrix}=\begin{pmatrix}1+2\lambda x\\ 1+2\lambda y\\ x^2+y^2-1\end{pmatrix}.$$

So the stationary points satisfy

$$x=y=-\frac{1}{2\lambda},\quad x^2+y^2=1.$$

This implies

$$\left(\frac{-1}{2\lambda}\right)^2+\left(\frac{-1}{2\lambda}\right)^2=1.$$

Thus $\lambda=\frac{\sqrt{2}}{2}$ or $\lambda=-\frac{\sqrt{2}}{2}$. And this further implies $(x^*,y^*)=(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})$ or $(x^*,y^*)=(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})$. However, simple calculation shows $(x^*,y^*)=(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})$ is a maximizer. So the minimizer is $(x^*,y^*)=(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})$.

For this kind of simple questions, we have a graphical interpretation. Consider the general problem

$$\begin{aligned} &\text{minimize } f(x,y), \\ &\text{subject to } g(x,y)=0, \end{aligned}$$

where $x,y\in\mathbb{R}$. The Lagrangian is

$$L(x,y,\lambda)=f(x,y)+\lambda g(x,y).$$

Then

$$\nabla L=0\Leftrightarrow\begin{cases}\nabla f(x,y)=-\lambda\nabla g(x,y)\\ g(x,y)=0.\end{cases}$$

The graphical interpretation is shown in Figure 1. We have several observations.

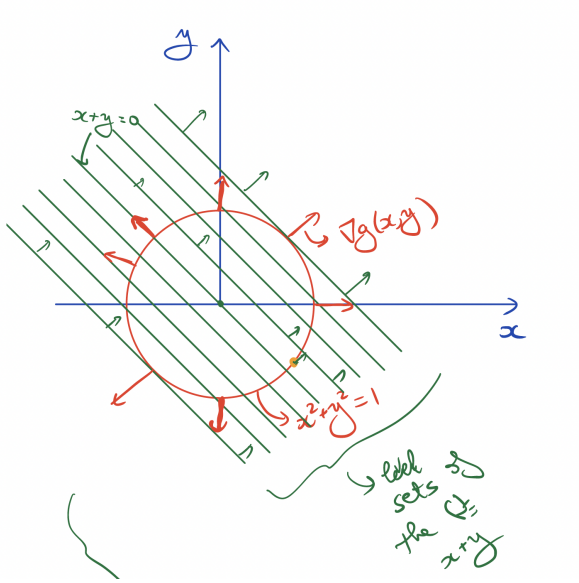

Figure 1: Graphical Interpretation

- At the stationary points of L, the red and green curves are tangent to each other.
- If the level sets associated with a particular point are not tangent to the constraint, then we can still move along the constraint (red curve) and increase/decrease the function $f(x, y)$.

In summary, at the stationary points of L, we have

- ∇f and ∇g are in the same direction. In other words, there is some λ such that $\nabla f = -\lambda \nabla g$.
- $g(x, y) = 0$.

Caveat: Stationary points of L may be a min, max or neither.

1.2 Multiple equality constraints

What if we have multiple equality constraints:

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to}$$
$$h_1(x) = 0,\ h_2(x) = 0,\ ...,\ h_p(x) = 0.$$

Then the Lagrangian is

$$L(x, v) = f(x) + \sum_{i=1}^{p} v_i h_i(x) = f(x) + v^\top h(x),$$

where $x \in \mathbb{R}^n$, $v \in \mathbb{R}^p$ and $h(x) = (h_1(x), ..., h_p(x))^\top$. Now

$$\nabla L = 0 \Leftrightarrow \begin{cases} \nabla f(x) = -\sum_{i=1}^{p} v_i \nabla h_i(x) \\ h_i(x) = 0 \text{ for } i = 1, 2, ..., p. \end{cases}$$

The graphical interpretation is that the intersection of the level surfaces $h_i(x) = 0$ is tangent to the level surface of f.

1.3 Inequality constraints

More generally, we can further involve inequality constraints:

$$\min_{x\in\mathbb{R}^n} f(x) \text{ subject to}$$
$$g_i(x) \leq 0 \text{ for } i = 1, ..., m,$$
$$h_i(x) = 0 \text{ for } i = 1, ..., p.$$

The Lagrangian associated with this is

$$L(x, \lambda, v) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} v_i h_i(x).$$

The variables λ_i and v_i are called Lagrange multipliers.

1.4 Examples of constrained optimization

Here we give some important examples.
Example. (Least squares)
We want to solve

$$\min_x x^\top x \text{ subject to}$$
$$Ax - b = 0.$$

The Lagrangian is $L(x, v) = x^\top x + v^\top (Ax - b)$. The reason is that $Ax - b = 0$ is equivalent to $\langle a_i, x\rangle - b_i = 0$, $i = 1, ..., p$. Then we can view $\langle a_i, x\rangle - b_i$ as $h_i(x)$. The Lagrangian is

$$L(x, v) = x^\top x - \sum_{i=1}^{p} v_i h_i(x) = x^\top x - \sum_{i=1}^{p} v_i(\langle a_i, x\rangle - b_i) = x^\top x + v^\top (Ax - b).$$

Example. (Linear optimization)
We want to solve

$$\min_x c^\top x \text{ subject to}$$
$$Ax = b,$$
$$x \geq 0.$$

The Lagrangian in this problem is $L(x, \lambda, v) = c^\top x - \lambda^\top x + v^\top (Ax - b)$.

Math 173B - Lecture 11: Lagrange Duality

1 Definitions and Examples

Definition 1.1. Given a Lagrangian $L(x, \lambda, \nu)$, its Lagrangian dual is defined to be $F(\lambda, \nu) := \inf_x L(x, \lambda, \nu)$.

Example 1. Dual of least square. Here, we have

$$\min_x x^\top x \text{ subject to}$$
$$Ax - b = 0.$$

The Lagrangian is $L(x, v) = x\top x + v^\top(Ax - b)$. So the dual is $F(\nu) = \min_x x\top x + v^\top(Ax - b)$. Since the optimizer is convex in x, we can find the minimizer by setting the gradient $\nabla_x L(x, v) = 0$, which is equivalent to $2x + A^\top \nu = 0$. Plugging in the optimizer $x^* = -\frac{A^\top \nu}{2}$ into $L(x, v)$ we get the dual of least square is $F(\nu) = -\frac{1}{4} v^\top A A^\top v - v^\top b$.

Example 2. Dual of linear Optimization. Here, we have

$$\min_x c^\top x \text{ subject to}$$
$$Ax = b,$$
$$x \geq 0.$$

The Lagrangian now is $L(x, \lambda, v) = c^\top x - \lambda^\top x + v^\top(Ax - b) = (c^\top + v^\top A - \lambda^\top)x - v^\top b$. To get its dual, we note that $v^\top(Ax - b)$ is a constant and $(c^\top + v^\top A - \lambda^\top)x$ is linear in x. Two cases can happen.

Either **Case I**, $c^\top + v^\top A - \lambda^\top = 0$ and we have $F(\lambda, \nu) = -v^\top b$. Or **Case II**, $c^\top + v^\top A - \lambda^\top \neq 0$ and we have $F(\lambda, \nu) = -\infty$.

Let $y := -\nu$, then when $\lambda \geq 0$, the dual of linear program is given by

$$F(\lambda, \nu) = \begin{cases} b^\top y & \text{when } A^\top y = c - \lambda \geq c, \\ -\infty & \text{otherwise.} \end{cases}$$

Definition 1.2. We define the following original constrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to}$$
$$g_i(x) \leq 0 \text{ for } i = 1, ..., m,$$
$$h_i(x) = 0 \text{ for } i = 1, ..., p.$$

as the primal optimization problem.

The Lagrangian associated with the above is

$$L(x, \lambda, v) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} v_i h_i(x).$$

The variables λ_i and v_i are called Lagrange multipliers.

Definition 1.3. We define the associated dual problem (w.r.t the above primal problem) to be

$$\max_{\lambda,\nu} F(\lambda,\nu) \quad \text{s.t. } \lambda \geq 0$$

where $F(\lambda,\nu) := \inf_x L(x,\lambda,\nu)$.

2 Duality Theory

Theorem 2.1 (Weak Duality)**.** *The weak duality theorem states that for any feasible primal solution $\alpha^* = f(x^*)$ and any feasible dual solution $\beta^* = F(\lambda^*,\nu^*)$, the optimal value of the dual problem provides a lower bound on the optimal value of the primal problem:*

$$\min f(x) = \alpha^* \geq \beta^* = \max_{\lambda \geq 0} F(\lambda,\nu). \tag{1}$$

Proof. We begin with the definition of the Lagrangian function:

$$L(x,\lambda,\nu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \nu_j h_j(x), \tag{2}$$

where $\lambda_i \geq 0$ for all i.

Since x is feasible for the primal problem, we have:

$$g_i(x) \leq 0, \quad h_j(x) = 0.$$

Multiplying both sides of the inequality constraints by their corresponding nonnegative multipliers λ_i and summing, we obtain:

$$\sum_i \lambda_i g_i(x) \leq 0.$$

Since the equality constraints hold exactly, their contribution vanishes:

$$\sum_j \nu_j h_j(x) = 0.$$

Thus, adding these terms to the primal objective function does not increase its value:

$$f(x) \geq L(x,\lambda,\nu).$$

By the definition of the dual function and taking minimum over x at both sides we have

$$F(\lambda,\nu) = \inf_x L(x,\lambda,\nu),$$

and

$$\alpha^* = \min f(x) \geq F(\lambda,\nu).$$

Since the above holds for all feasible dual variables, we can take the maximum over dual variables and obtain the weak duality result:

$$\alpha^* = \min f(x) \geq \max F(\lambda^*,\nu^*) = \beta^*.$$

This completes the proof of weak duality.

□

2.1 Weak and Strong Duality

Weak duality always holds. For feasible x and (λ, ν):

$$\alpha^* = \min f(x) \geq \max F(\lambda^*, \nu^*) = \beta^*.$$

Strong duality needs extra assumptions.

Theorem 2.2 (Strong Duality)**.** *If Slater's condition holds, we have*

$$\alpha^* = \min f(x) = \max F(\lambda^*, \nu^*) = \beta^*. \tag{3}$$

Strong duality will allow us to solve the primal problem by finding the max of the dual problem.

Definition 2.3 (Slater's Condition)**.** A primal optimization problem satisfies Slater's condition if

- f is convex
- All $g_i(x)$ are convex
- All $h_j(x)$ are linear
- $\exists\, \bar{x}$ s.t. $g_i(\bar{x}) < 0 \quad \forall\, i = 1, ..., m,\ h_j(\bar{x}) = 0 \quad \forall\, j = 1, ..., p.$

Example 3. When the objective function f and all the constraints $g_i(x), h_j(x)$ are linear, strong duality hold and we have $\alpha^* = \beta^*$

Math 173B - Lecture 12: Detour: Classification Problems, Support Vector Machine and Duality

1 Classification

Somebody hands you accurately labeled data

$$\{(x_1, y_1), ..., (x_n, y_n)\},$$

where $x_i \in \mathbb{R}^d$ and y_i are in the label set such as $\{+1, -1\}$. So the x_i's are data points and y_i's are their labels. The goal is that given the data points and labels, learn a function $f : \mathbb{R}^d \to \mathbb{R}$ so that given a new data point x with unknown label y, we have $f(x) > 0$ when $y = +1$ and $f(x) < 0$ when $y = -1$.

1.1 Linear Discrimination

To simplify matters, we assume that f is an affine function

$$f(x) = a^\top x + b,$$

where $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are unknown. So we want

$$\begin{cases} a^\top x_i + b > 0 & \text{when } y_i > 0 \\ a^\top x_i + b < 0 & \text{when } y_i < 0. \end{cases} \tag{1}$$

In other words, we want t a hyperplane that separates the clusters. (See Figure 1)

1.2 Support Vector Machine

Instead of (1), it will be more convenient to deal with

$$\begin{cases} a^\top x_i + b > +1 & \text{when } y_i > 0 \\ a^\top x_i + b < -1 & \text{when } y_i < 0. \end{cases} \tag{2}$$

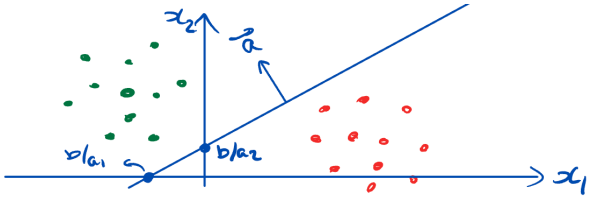

Figure 1: The classification problem

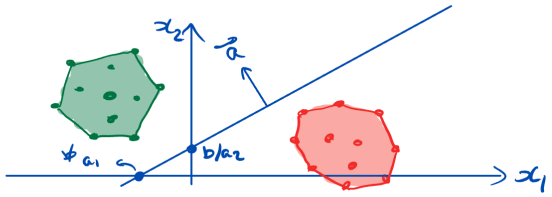

Figure 2: Graphical illustration of Proposition 1

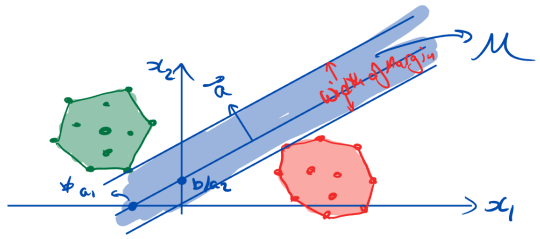

Figure 3: Graphical illustration of Definition 1.1

It is obvious that (1) is equivalent to (2). And (2) is further equivalent to

$$\begin{cases} a^\top x_i + b > +1 & \text{when } y_i > 0 \\ a^\top(-x_i) + (-b) > +1 & \text{when } y_i < 0. \end{cases} \tag{3}$$

Now we let

$$I = \{i : y_i = +1\},$$
$$J = \{i : y_i = -1\}.$$

Proposition 1. *(3) is feasible if and only if the convex hulls of $A = \{x_i : i \in I\}$ and $B = \{x_i : i \in J\}$ do not intersect.*

The graphical interpretation of Proposition 1 is illustrated in Figure 2. We assume that the convex hulls of A and B don't intersect so there are choices of $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that make (3) hold.

Now our goal is to find the "best" choice of $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Definition 1.1. The set $M = \{x \in \mathbb{R}^d,\ -1 \le a^\top x + b \le +1\}$ is called the classification margin and contains no data points.

The graphical interpretation of Definition 1.1 is illustrated in Figure 3. The width of M is the distance between the hyperplanes:

$$\{x : a^\top x + b = 1\} \text{ and } \{x : a^\top x + b = -1\}.$$

Fact: Width of M is $\frac{2}{\|a\|}$. (left as exercise)

So in summary, we want to solve

$$\max_{a,b} \frac{2}{\|a\|} \text{ subject to } \begin{cases} a^\top x_i + b \ge 1, & \forall i \in I, \\ a^\top x_i + b \le -1, & \forall i \in J. \end{cases}$$

The above problem is equivalent to

$$\min_{a,b} \frac{\|a\|^2}{4} \text{ subject to } \begin{cases} a^\top x_i + b \geq 1, & \forall i \in I, \\ a^\top x_i + b \leq -1, & \forall i \in J. \end{cases}$$

We call the above equivalent problem **the primal problem**, denoted as (P).

Next we introduce the dual of (P). Here are some steps we need.

- Compute Lagrangian function $L(a, b, \lambda, \mu)$.
- Compute the dual function

$$F(\lambda, \mu) = \min_{a,b} L(a, b, \lambda, \mu).$$

- Write the dual optimization problem.

$$\max_{\lambda,\mu} F(\lambda, \mu) \text{ subject to } \lambda \geq 0 \text{ and } \mu \geq 0.$$

First we can compute the Lagrangian

$$L(a, b, \lambda, \mu) = \frac{\|a\|^2}{4} + \sum_{i \in I} \lambda_i (1 - a^\top x_i - b) + \sum_{j \in J} \mu_j (b + a^\top x_j + 1),$$

where λ_i's and μ_i's are the dual variables. Now we rearrange the Lagrangian:

$$L(a, b, \lambda, \mu) = \frac{\|a\|^2}{4} + \langle a, \sum_{j \in J} \mu_j x_j - \sum_{i \in I} \lambda_i x_i \rangle + \langle b, \sum_{j \in J} \mu_j - \sum_{i \in I} \lambda_i \rangle + \sum_{i \in I} \lambda_i + \sum_{j \in J} \mu_j.$$

Now we know $F(\lambda, \mu) = \min_{a,b} L(a, b, \lambda, \mu)$. Since L is convex in (a, b), then setting and solving

$$\nabla_a L = 0 \text{ and } \nabla_b L = 0$$

yields the minimizer. So

$$\begin{cases} \frac{a}{2} + \sum_{j \in J} \mu_j x_j - \sum_{i \in I} \lambda_i x_i = 0, \\ \sum_{j \in J} \mu_j - \sum_{i \in I} \lambda_i = 0. \end{cases} \tag{4}$$

Here, the second equality is needed to prevent $\langle b, \sum_{j \in J} \mu_j - \sum_{i \in I} \lambda_i \rangle$ from going to infinity. Now we can plug (4) into the expression of L to get

$$F(\lambda, \mu) = -\| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} \mu_j x_j \|^2 + \sum_{i \in I} \lambda_i + \sum_{j \in J} \mu_j.$$

Then the dual problem reads

$$\max_{\lambda,\mu} \left(-\| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} \mu_j x_j \|^2 + \sum_{i \in I} \lambda_i + \sum_{j \in J} \mu_j \right) \text{ subject to } \begin{cases} \lambda \geq 0, \\ \mu \geq 0, \\ \sum_{i \in I} \lambda_i = \sum_{j \in J} \mu_j. \end{cases}$$

Math 173B - Lecture 13: Dual Gaps, Complementary Slackness, KKT Conditions

1 Lecture 12 (continue)

Recall the dual problem reads

$$\max_{\lambda,\mu}\left(-\|\sum_{i\in I}\lambda_i x_i-\sum_{j\in J}\mu_j x_j\|^2+\sum_{i\in I}\lambda_i+\sum_{j\in J}\mu_j\right) \text{ subject to } \begin{cases} \lambda\geq 0,\\ \mu\geq 0,\\ \sum_{i\in I}\lambda_i=\sum_{j\in J}\mu_j. \end{cases}$$

Now let us manipulate this a bit by setting

$$\sum_{i\in I}\lambda_i=s\geq 0,$$

and setting $\lambda_i^{'}=\lambda_i/s, \mu_i^{'}=\mu_i/s$. Then the above problem becomes

$$\max_{\lambda^{'},\mu^{'}}\left(-s^2\|\sum_{i\in I}\lambda_i^{'}x_i-\sum_{j\in J}\mu_j^{'}x_j\|^2+2s\right) \text{ subject to } \begin{cases} \lambda^{'}\geq 0,\\ \mu^{'}\geq 0,\\ s\geq 0,\\ \sum_{i\in I}\lambda_i^{'}=\sum_{j\in J}\mu_j^{'}=1. \end{cases}$$

Notice that the maximum in the variable s is attained when

$$s^*=\frac{1}{\|\sum_{i\in I}\lambda_i^{'}x_i-\sum_{j\in J}\mu_j^{'}x_j\|^2}.$$

So that the dual now is (by substituting s^* above):

$$\max_{\lambda^{'},\mu^{'}}\frac{1}{\|\sum_{i\in I}\lambda_i^{'}x_i-\sum_{j\in J}\mu_j^{'}x_j\|^2} \text{ subject to } \begin{cases} \lambda^{'}\geq 0,\\ \mu^{'}\geq 0,\\ \sum_{i\in I}\lambda_i^{'}=\sum_{j\in J}\mu_j^{'}=1. \end{cases}$$

A further rewriting gives

$$\min_{\lambda,\mu}\|\sum_{i\in I}\lambda_i x_i-\sum_{j\in J}\mu_j x_j\|^2 \text{ subject to } \begin{cases} \lambda\geq 0,\\ \mu\geq 0,\\ \sum_{i\in I}\lambda_i=\sum_{j\in J}\mu_j=1. \end{cases} \tag{1}$$

This is the **dual problem** (D) we finally get. Before we give a geometric interpretation of the dual, let us check whether we have strong duality.

Recall the primal problem:

$$\min_{a,b}\frac{\|a\|^2}{4} \text{ subject to } \begin{cases} a^\top x_i+b\geq 1, & \forall i\in I,\\ a^\top x_i+b\leq -1, & \forall i\in J. \end{cases}$$

Here we can verify the Slater's condition for it (verify it by yourself):

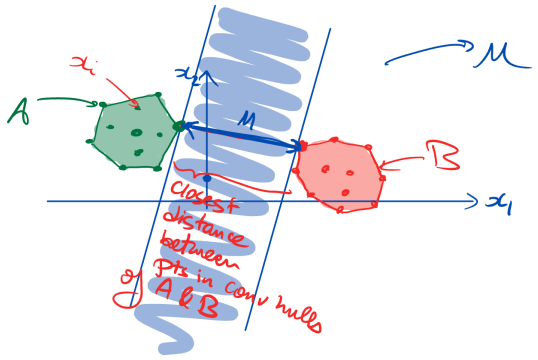

Figure 1: Graphical illustration of the dual problem

- f is convex. (why?)
- g_i's are convex ($\langle a_i, x\rangle + b - 1$ is a convex function)
- h_j's are linear. (there are no h_j's)
- There exist $\bar{a}$ and $\bar{b}$ so that $g_i(\bar{a}, \bar{b}) < 0$ for all i. (why? see the below remark)

Remark 1. Only the last one is a bit non-trivial. Here are the hints. Notice the last point is the same as saying that the inequalities

$$\langle \bar{a}, x_i\rangle + \bar{b} - 1 > 0 \text{ for all } i \in I,$$
$$\langle \bar{a}, x_j\rangle + \bar{b} + 1 < 0 \text{ for all } j \in J,$$

can be satisfied strictly by some points $\bar{a}$ and $\bar{b}$. But we assumed that the convex hulls of x_i, where $i \in I$ and x_j, where $j \in J$ do not intersect. So we are fine. (why? think about it)

Now since Slater's condition holds, we have strong duality. In other words, the optimal value of the original problem

$$\max_{a,b} \frac{2}{\|a\|} \text{ subject to } \begin{cases} a^\top x_i + b \geq 1, & \forall i \in I, \\ a^\top x_i + b \leq -1, & \forall i \in J. \end{cases}$$

is the same as the optimal value of the dual problem

$$\min_{\lambda,\mu} \|\sum_{i\in I} \lambda_i x_i - \sum_{j\in J} \mu_j x_j\|^2 \text{ subject to } \begin{cases} \lambda \geq 0, \\ \mu \geq 0, \\ \sum_{i\in I} \lambda_i = \sum_{j\in J} \mu_j = 1. \end{cases}$$

The geometric interpretation of the primal problem is maximizing the width of the classification margin. But what is the geometric interpretation of the dual problem above? Noting that $\sum_{i\in I} \lambda_i x_i$ with $\sum_{i\in I} \lambda_i$ and $\lambda_i \geq 0$ is a point in the convex hull of the set $\{x_i, i \in I\}$ and same for μ_j's. We see that the dual problem finds two points in the convex hulls of our sets A and B that are closest. This geometric interpretation is illustrated in Figure 1.

Strong duality tells us this distance is the same as the margin associated with the best separating hyperplane.

2 Dual Gaps, Complementary Slackness, KKT Conditions

Recall that in general, we want to solve

$$\min_x f(x) \text{ subject to } \begin{cases} g_i(x) \leq 0, \text{ for } i = 1, ..., m \\ h_j(x) = 0, \text{ for } j = 1, ..., p. \end{cases}$$

The Lagrangian of this problem is

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x).$$

And the Lagrangian dual is

$$F(\lambda, \nu) = \min_{x \in D} L(x, \lambda, \nu),$$

which finally leads to the dual problem:

$$\max_{\lambda, \nu} F(\lambda, \nu) \text{ subject to } \lambda \geq 0.$$

Recall that the weak duality always holds, i.e., the maximum of the dual problem (denoted as β^*) is less than or equal to the minimum of the primal problem (denoted as α^*). Namely, $\beta^* \leq \alpha^*$. While strong duality, i.e., $\beta^* = \alpha^*$, does not always hold, but it sometimes hold (for example, Slater's condition is satisfied).

2.1 Dual Gap

Sometimes it is easier to solve dual than primal, and moreover, we always have $\beta^* \leq \alpha^*$. This means that for $\forall x, \lambda, \nu$, we have

$$F(\lambda, \nu) \leq \beta^* \leq \alpha^* \leq f(x),$$

which implies

$$f(x) - \alpha^* \leq f(x) - F(\lambda, \nu).$$

So if $f(x) - F(\lambda, \nu)$ is small then we know that $f(x) - \alpha^*$ is small, which would mean that x is close to optimal.

Definition 2.1. For an optimization problem with objective f, dual F, and given a primal feasible point x, a dual feasible point (λ, ν), the quantity

$$f(x) - F(\lambda, \nu)$$

is defined as the duality gap.

Now we introduce a potential usefulness of this definition. If you have an algorithm that produces a sequence of primal and dual feasible points:

$$x^{(i)} \text{ and } (\lambda^{(i)}, \nu^{(i)}),$$

then if $f(x^{(i)}) - F(\lambda^{(i)}, \nu^{(i)}) \leq \epsilon$, you know that $f(x^{(i)}) - \alpha^* \leq \epsilon$.

Math 173B - Lecture 14: Complementary Slackness and KKT Conditions

1 Complementary Slackness

Lemma 1.1. *Let x^* be a primal optimal solution, and let (λ^*, ν^*) be dual optimal solution, and suppose we have strong duality. Then the complementary slackness conditions hold:*

1. *x^* minimizes $L(x, \lambda^*, \nu^*)$*
2. *$\lambda_i^* g_i(x^*) = 0, \quad \forall i.$*

This means that if an inequality constraint $g_i(x) \leq 0$ is not active (i.e., $g_i(x^*) < 0$), then the corresponding Lagrange multiplier must be zero: $\lambda_i^* = 0$.

Proof. By strong duality, the optimal primal and dual solutions satisfy:

$$f(x^*) = \max_{\lambda \geq 0, \nu} F(\lambda, \nu) = F(\lambda^*, \nu^*) = \inf_x L(x, \lambda^*, \nu^*),$$

where the Lagrangian is defined as:

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x).$$

Notice the simple fact

$$\inf_x L(x, \lambda^*, \nu^*) \leq L(x^*, \lambda^*, \nu^*).$$

We now have:

$$f(x^*) \leq f(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i(x^*) + \sum_{j=1}^{p} \nu_j^* h_j(x^*),$$

which implies

$$0 \leq \sum_{i=1}^{m} \lambda_i^* g_i(x^*) + \sum_{j=1}^{p} \nu_j^* h_j(x^*)$$

Furthermore, since x^* and λ^* are feasible, we have:

$$\lambda_i^* \geq 0, \quad g_i(x^*) \leq 0, \quad h_j(x^*) = 0.$$

This implies that all the inequalities in the feasibility conditions must hold as equalities, leading to:

$$x^* = \arg\min_x L(x, \lambda^*, \nu^*),$$

and

$$\sum_{i=1}^{m} \lambda_i^* g_i(x^*) = 0.$$

Since each $\lambda_i^* \geq 0$ and $g_i(x^*) \leq 0$, it follows that whenever $g_i(x^*) < 0$, we must have $\lambda_i^* = 0$; otherwise, their product would be negative, violating the equation. □

2 KKT Conditions

We still consider our standard optimization problem:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0, \quad i = 1, ..., m, \\ & h_j(x) = 0, \quad j = 1, ..., p. \end{aligned}$$

Definition 2.1. We say primal variable x^* and dual variable λ^*, ν^* satisfy the KKT conditions if:

1. Primal inequality feasibility: $g_i(x^*) \le 0, \forall i = 1, ..., m$
2. Primal inequality feasibility: $h_j(x^*) = 0, \forall j = 1, ..., p$
3. Dual feasibility: $\lambda_i^* \ge 0, \forall i = 1, ..., m$
4. Complementary slackness: $\lambda_i^* g_i(x^*) = 0, \forall i = 1, ..., m$
5. Stationary Condition: $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(x^*) = 0$.

2.1 Necessary Condition

Theorem 2.2. *Suppose x^* is primal opt, λ^*, ν^* is dual opt and strong duality holds. Then we have x^*, λ^*, ν^* satisfy the KKT conditions.*

Proof. Feasibility automatically holds by assumption. We just proved complementary slackness by last lemma. The only thing left to show is (5), stationary condition.

From the previous lemma, we know that the primal opt x^* minimizes the Lagrangian $L(x, \lambda^*, \nu^*)$:

$$L(x, \lambda^*, \nu^*) = f(x) + \sum_{i=1}^m \lambda_i^* g_i(x) + \sum_{j=1}^p \nu_j^* h_j(x).$$

By the first order necessary condition for a local minimizer and taking gradient of $L(x, \lambda^*, \nu^*)$ w.r.t x, we get:

$$\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(x^*) = 0.$$

□

2.2 Sufficient Condition

Theorem 2.3. *Suppose f is convex, all g_i's are convex, and all h_j's are affine. Suppose there exists x^*, λ^*, ν^* satisfying the KKT conditions. Then we can conclude that strong duality holds.*

Proof. By KKT conditions, we have x^* is primal feasible and λ^*, ν^* are dual feasible. Since f is convex, all g_i's are convex, and all h_j's are affine, we have

$$L(x, \lambda^*, \nu^*) = f(x) + \sum_{i=1}^m \lambda_i^* g_i(x) + \sum_{j=1}^p \nu_j^* h_j(x)$$

is convex in x. This implies any stationary point is a local minimizer, thus a global minimizer. By the stationary condition: $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$, meaning x^* minimizes $L(x, \lambda^*, \nu^*)$. By definition of the Lagrangian dual we have:

$$F(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) = f(x^*) + \sum_{i=1}^m \lambda_i^* g_i(x^*) + \sum_{j=1}^p \nu_j^* h_j(x^*) = f(x^*),$$

where we used complementary slackness (4) and dual feasibility condition (2) in the last equal sign.

By weak duality, for any given feasible x^*:

$$f(x^*) \geq F(\lambda^*, \nu^*).$$

Thus we know λ^*, ν^* maximized F and strong duality holds. $\square$

Math 173B - Lecture 15: Dual Ascent, Augmented Lagrangian and Method of Multipliers

After our discussion on duality theory, let us focus on solving problems of the type,

$$\min_{x\in\mathbb{R}^n} f(x) \text{ subject to } Ax = b,$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$. This is the primal problem (P). Specifically, we want algorithms for solving (P).

1 Dual Ascent

1.1 The algorithm

The idea is to consider the dual problem. The Lagrangian is

$$L(x, y) = f(x) + y^\top (Ax - b),$$

where $y \in \mathbb{R}^m$. The dual is

$$F(y) = \inf_x L(x, y) = f(x_y) + y^\top (Ax_y - b).$$

Here we suppose x_y achieves the infimum (minimum) for a fixed y. And the dual optimization problem (D) is

$$\max_y F(y).$$

If we have strong duality, then

$$\max_y F(y) = \left(\min_x f(x) \text{ subject to } Ax = b\right).$$

Moreover, if y^* solves the dual problem, an optimizer for the primal (P) can be found by

$$x^* = \arg\min_x L(x, y^*).$$

(Why? think about it.).

Next, we introduce the dual descent algorithm. The general procedure is: for a given y, solve:

$$x^\dagger = \arg\min_x L(x, y),$$

and then update y by trying to maximize the dual. To be more precise:

- $x^{(t+1)} = \arg\min_x L(x, y^{(t)})$.
- $y^{(t+1)} = y^{(t)} + \alpha_t \underbrace{(Ax^{(t+1)} - b)}_{\nabla_y L}$.

Here, the second step can be interpreted as a gradient ascent step on the dual variable y with step size α_t.

Theorem 1.1 ("Convergence guarantee"). *If α_t is chosen correctly and the function is nice (convex, ...) then*

$$x^{(t+1)} \to x^*,$$
$$y^{(t+1)} \to y^*.$$

1.2 Pros of Dual Ascent

Suppose that we could write

$$f(x) = \sum_{i=1}^{n} f_i(x_i).$$

(For example $f(x) = \sum_{i=1}^{n} \|x_i\|^2$.)
Then

$$L(x, y) = \left[\sum_{i=1}^{n} f_i(x_i)\right] + \left[y^\top (Ax - b)\right].$$

We could note that $Ax = \sum_{i=1}^{n} a_i x_i$, where a_i is the i-th column of A. So

$$L(x, y) = \sum_{i=1}^{n} \left[f_i(x_i)\right] + \sum_{i=1}^{n} \left[y^\top a_i x_i - \frac{y^\top b}{n}\right] = \sum_{i=1}^{n} L_i(x_i, y),$$

where

$$L_i(x, y) = f_i(x) + y^\top a_i x_i - \frac{y^\top b}{n}.$$

Then we can solve for each variable separately (say, in parallel) and we can write

- $x_i^{(t+1)} = \arg\min_{x_i} L_i(x_i, y^{(t)})$. (in parallel for each $i = 1, ..., n$)
- $y^{(t+1)} = y^{(t)} + \alpha_t (Ax^{(t+1)} - b)$.

This specific version is called the dual decomposition method.

2 Augmented Lagrangian and method of multipliers

The purpose of the section is to introduce methods that are developed to robustify dual ascent and weaken some of the assumptions needed for it to work.

2.1 The Algorithm

Definition 2.1. The function

$$L_\rho(x, y) = f(x) + y^\top (Ax - b) + \frac{\rho}{2} \|Ax - b\|^2$$

with $\rho \geq 0$ (called the penalty parameter) is known as the augmented Lagrangian for (P).

Remark 1. It is easy to notice when $\rho = 0$, $L_0(x, y) = L(x, y)$.

Now we introduce the method of multipliers (MM):

- $x^{(t+1)} = \arg\min_x L_\rho(x, y^{(t)})$.
- $y^{(t+1)} = y^{(t)} + \rho(Ax^{(t+1)} - b)$.

Remark 2. Notice that in the second step of the update, the step size is set to ρ, which is the penalty parameter, instead of an arbitrary α.

Remark 3. MM converges under more general conditions than dual ascent. (e.g. f can take $+\infty$ as a value and f need not be strictly convex.)

Why do we use ρ as a step size? Recall the primal (P):

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

with optimality conditions:

$$Ax^* - b = 0, \tag{1}$$
$$\nabla f(x^*) + A^\top y^* = 0. \tag{2}$$

In the first step of MM, we have

$$x^{(t+1)} = \arg\min_x L_\rho(x, y^{(t)}).$$

So to find $x^{(t+1)}$, we have to solve

$$\nabla_x L_\rho(x, y^{(t)}) = 0,$$

which is equivalent to

$$0 = \nabla f(x^{(t+1)}) + A^\top \underbrace{(y^{(t)} + \rho(Ax^{(t+1)} - b))}_{y^{(t+1)}}.$$

In other words,

$$0 = \nabla f(x^{(t+1)}) + A^\top y^{(t+1)}.$$

If we compare the above identity to (2), we notice $(x^{(t+1)}, y^{(t+1)})$ always satisfies (2). At the same time, we have

$$Ax^{(t)} - b \to 0$$

as $t \to 0$, giving primal feasibility (1).

2.2 Downside of MM

Even when f is separable, i.e.,

$$f(x) = \sum_{i=1}^n f_i(x_i),$$

$L_\rho(x, y)$ is not separable in general. This is because

$$\|Ax - b\|^2 = x^\top A^\top Ax - 2x^\top A^\top b + b^\top b.$$

The first term $x^\top A^\top Ax$ is a stumbling block because it mixes all the coefficients of x_i. For example,

$$A = (1, 1),$$
$$A^\top A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Thus $x^\top A^\top Ax = x_1^2 + 2x_1x_2 + x_2^2$. This implies we can not parallelize the method of multipliers easily. The solution is to develop a new algorithm, which is called ADMM. We will introduce ADMM next time.

Math 173B - Lecture 16: Alternating Direction Method of Multipliers (ADMM)

1 Alternating Direction Method of Multipliers (ADMM)

In this section, we will introduce a method that has the decomposability of dual descent and superior convergence properties of the method of multipliers.

Suppose we consider the problem:

$$\min_{x,z} f(x) + g(z) \text{ subject to } Ax + Bz = c,$$

where $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$, and $c \in \mathbb{R}^p$.

An example of this kind of problems is

$$\min_{u} \sum_{i=1}^{n+m} u_i^2 \text{ subject to } Mu = c,$$

where $M \in \mathbb{R}^{p\times(n+m)}$ and $c \in \mathbb{R}^p$. This problem can be rewritten as

$$\min_{x,z} \|x\|^2 + \|z\|^2 \text{ subject to } M_1x + M_2z = c,$$

where $x = (u_1, ..., u_n) \in \mathbb{R}^n$, $z = (u_{n+1}, ..., u_{n+m}) \in \mathbb{R}^m$, $M_1 \in \mathbb{R}^{p\times n}$ is the matrix consisting of the first n columns of M and $M_2 \in \mathbb{R}^{p\times m}$ is the matrix consisting of the remaining columns of M. (why they are equivalent? think about it)

Now suppose

$$\alpha^* = \inf\{f(x) + g(z) \mid Ax + Bz = c\}.$$

From the augmented Lagrangian:

$$L_\rho(x, z, y) = f(x) + g(z) + y^\top(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2.$$

Then the ADMM algorithm is

- $x^{(t+1)} = \text{argmin}_x L_\rho(x, z^{(t)}, y^{(t)})$.
- $z^{(t+1)} = \text{argmin}_z L_\rho(x^{(t+1)}, z, y^{(t)})$.
- $y^{(t+1)} = y^{(t)} + \rho(Ax^{(t+1)} + Bz^{(t+1)} - c)$.

Remark 1. In this case, the method of multipliers would have given

- $(x^{(t+1)}, z^{(t+1)}) = \text{argmin}_{(x,z)} L_\rho(x, z, y^{(t)})$.
- $y^{(t+1)} = y^{(t)} + \rho(Ax^{(t+1)} + Bz^{(t+1)} - c)$.

(x, z) are jointly optimized at each step of MM while they are separate in ADMM.

Now we introduce the scaled form of ADMM. Define the residual r, at (x, z) as $r = Ax + Bz - c$. Then we know

$$\begin{aligned}
&y^\top (Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2 \\
&= y^\top r + \frac{\rho}{2}\|r\|^2 \\
&= \frac{\rho}{2}\|r + \frac{1}{\rho}y\|^2 - \frac{1}{2\rho}\|y\|^2 \\
&= \frac{\rho}{2}\|r + u\|^2 - \frac{\rho}{2}\|u\|^2,
\end{aligned}$$

where $y = \rho u$. Since the term $y^\top (Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2$ appears in the augmented Lagrangian we can write an equivalent form of ADMM:

- $x^{(t+1)} = \arg\min_x (f(x) + \frac{\rho}{2}\|Ax + Bz^{(t)} - c + u^{(t)}\|^2)$.
- $z^{(t+1)} = \arg\min_z (g(z) + \frac{\rho}{2}\|Ax^{(t+1)} + Bz - c + u^{(t)}\|^2)$.
- $u^{(t+1)} = u^{(t)} + \underbrace{Ax^{(t+1)} + Bz^{(t+1)} - c}_{r^{(t+1)}}$.

Notice that $u^{(t)} = u^{(t-1)} + r^{(t)}$, so $u^{(t)} = u^{(0)} + \sum_{i=1}^{t-1} r^{(i)}$.
Now let us talk about the convergence. Assume the following,

- $\text{epi}(f) = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \le t\}$ is closed, convex, nonempty and so is $\text{epi}(g)$.
- The Lagrangian ($L(x, z, y)$) has a saddle point.

Then those two assumptions imply

- (Feasibility, asymptotically) $r^{(t)} \to 0$ as $t \to \infty$.
- $f(x^{(t)}) + g(z^{(t)}) \to \alpha^*$ as $t \to \infty$.
- (Dual variable converges to optimal) $y^{(t)} \to y^*$ as $t \to \infty$.

In practice, we always observe:

- Quickly makes initial progress.
- But can be slow to converge to very high accuracy.

Math 173B - Lecture 17: ADMM Continued

1 Alternating Direction Method of Multipliers (ADMM)

Recall we used ADMM to solve

$$\min_{x,z} f(x) + g(z) \text{ subject to } Ax + Bz = c,$$

where $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$, and $c \in \mathbb{R}^p$.

From the augmented Lagrangian

$$L_\rho(x, z, y) = f(x) + g(z) + y^\top (Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2,$$

we can derive the ADMM algorithm as follows

$$\begin{cases} x^{(t+1)} & = \arg\min_x L_\rho(x, z^{(t)}, y^{(t)}), \\ z^{(t+1)} & = \arg\min_z L_\rho(x^{(t+1)}, z, y^{(t)}), \\ y^{(t+1)} & = y^{(t)} + \rho(Ax^{(t+1)} + Bz^{(t+1)} - c). \end{cases}$$

2 Variations and Practical Considerations

In practice, one may choose ρ to vary with iterations, i.e., use a step size ρ_t that changes over time. A common strategy is to check whether $r^{(t+1)} = Ax^{(t+1)} + Bz^{(t+1)} - c$ is large or small (compared to dual feasibility):

- Increase ρ_t when $\|r^{(t+1)}\|$ is large.
- Decrease ρ_t when $\|r^{(t+1)}\|$ is small.
- keep ρ_t unchanged otherwise

The idea is that ρ places a penalty on violating primal feasibility, so we adjust it dynamically.

In practice, it may be difficult to find a closed-form solution for

$$x^{(t+1)} = \arg\min_x L_\rho(x, z^{(t)}, y^{(t)}),$$

so we can approximate it using gradient descent (GD) or conjugate gradient (CG) methods.

3 Examples of Problems Where ADMM Can Be Applied

Recall that the x-update step of ADMM is:

$$x^{(t+1)} = \arg\min_x \left(f(x) + \frac{\rho}{2}\|Ax + (Bz^{(t)} - c + u^{(t)})\|^2 \right),$$

using the scaled form of ADMM:

- $x^{(t+1)} = \arg\min_x (f(x) + \frac{\rho}{2}\|Ax + Bz^{(t)} - c + u^{(t)}\|^2)$.
- $z^{(t+1)} = \arg\min_z (g(z) + \frac{\rho}{2}\|Ax^{(t+1)} + Bz - c + u^{(t)}\|^2)$.
- $u^{(t+1)} = u^{(t)} + \underbrace{Ax^{(t+1)} + Bz^{(t+1)} - c}_{r^{(t+1)}}$.

Notice that $u^{(t)} = u^{(t-1)} + r^{(t)}$, so $u^{(t)} = u^{(0)} + \sum_{i=1}^{t-1} r^{(i)}$.

3.1 Proximity Operator

To simply notation, we will just drop the step index t and use x^+ as the updated x. Consider the case where $A = I$, then the update simplifies to:

$$x^+ = \text{prox}_{f,\rho}(v),$$

where $\text{prox}_f = \arg\min_x \left(f(x) + \frac{\rho}{2}\|x - v\|^2\right)$ is the proximity operator and $v = -Bz + c - u$

If $f(x)$ is the indicator function of a non-empty, closed, convex set C, i.e.,

$$f(x) = \mathbb{1}_C(x) = \begin{cases} 0, & x \in C, \\ \infty, & x \notin C, \end{cases}$$

then the update simplifies to the projection onto C:

$$x^+ = \Pi_C(v).$$

Exercise: Verify this by yourself.

For example, if C is the non-negative orthant $\{x \mid x_i \geq 0\}$, then the update is:

$$x^{(t+1)} = \max(v, 0) = (v)_+ = \text{ReLU}(v).$$

3.2 Quadratic Objectives

Another important setting is when f is a convex quadratic. If $f(x) = \frac{1}{2}x^T Px + q^T x + r$ with P symmetric positive semi-definite, then the update step is given by (assuming $P + \rho A^T A$ invertible):

$$x^+ = (P + \rho A^T A)^{-1}(\rho A^T v - q).$$

Exercise: Verify this by yourself.

Hint: Finding x^+ is equivalent to solving a linear system in this case:

$$(P + \rho A^T A)x^+ = \rho A^T v - q,$$

which can be efficiently solved using direct solvers or iterative methods like conjugate gradient.

- Smooth objectives, such as those encountered in practical optimization problems, may be handled using first-order methods like gradient descent (GD) or quasi-Newton methods for improved convergence.
- A practical trick often used in practice is to warm-start the iterates $x^{(t+1)}$ so that the optimization algorithm for solving

$$x^{(t+1)} = \arg\min_x \left(f(x) + \frac{\rho}{2}\|Ax - v^{(t)}\|^2\right)$$

is initialized with the previous iterate $x^{(t)}$, improving efficiency.

4 Constrained Convex Optimization with ADMM

So far, we considered the problem:

$$\min_{x,z} f(x) + g(z) \text{ subject to } Ax + Bz = c,$$

and developed ADMM methods for it. More generally, we can rewrite an optimization problems

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

into ADMM form:

$$\min_{x,z} f(x) + g(z) \text{ subject to } x - z = 0,$$

where $g(z) = \mathbb{1}_C(z)$.

Exercise: Verify the equivalence by yourself

The augmented Lagrangian (scaled form) is then given by:

$$L_\rho(x, z, u) = f(x) + \mathbb{1}_C(z) + \frac{\rho}{2}\|x - z + u\|^2.$$

The ADMM algorithm in scaled form is:

$$\begin{aligned}
x^{(t+1)} &= \arg\min_x \left(f(x) + \frac{\rho}{2}\|x - z^{(t)} + u^{(t)}\|^2 \right), \\
z^{(t+1)} &= \Pi_C(x^{(t+1)} + u^{(t)}), \\
u^{(t+1)} &= u^{(t)} + x^{(t+1)} - z^{(t+1)}.
\end{aligned}$$

The z-update often involves projecting onto a convex set, while the x-update involves the proximal operator we've seen earlier.

Math 173B - Lecture 18: Some examples of optimization problems

1 The problem of finding intersections of closed convex sets

Suppose we want to find a point in the intersection of two closed convex sets: C and D.

We first introduce a classical approach that goes back to Von Neuman in the 1930's. This method is called *Alternating Pojections*:

- $x^{(t+1)} = \Pi_C(z^{(t)})$.
- $z^{(t+1)} = \Pi_D(x^{(t+1)})$

See Figure 1 for geometric interpretation.

As an alternative approach, we can use ADMM. The first step is to rewrite the problem that fits ADMM. To be more specific, the problem should be in the form of

$$\min_{x,z} f(x) + g(z) \text{ subject to } Ax + Bz = c.$$

In this case, we can take $f(x)$ to be the indicator function of C: $\mathbb{1}_C(x)$, $g(x)$ to be the indicator function of D: $\mathbb{1}_D(z)$, A, B to be I and $-I$, and c to be 0.

Writing the augmented Lagrangian and working through the math, we get the ADMM in scaled form:

- $x^{(t+1)} = \Pi_C(z^{(t)} - u^{(t)})$.
- $z^{(t+1)} = \Pi_D(x^{(t+1)} + u^{(t)})$.
- $u^{(t+1)} = u^{(t)} + x^{(t+1)} - z^{(t+1)}$

This is also called Dykstra's alternating projections method, and it is more efficient than the classical method mentioned above.

This can further be generalized to an algorithm for finding a point in the intersection of N closed, convex sets.

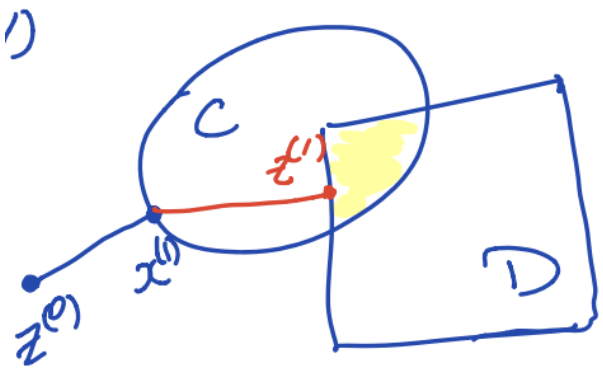

Figure 1: Graphical illustration of Alternating Projections

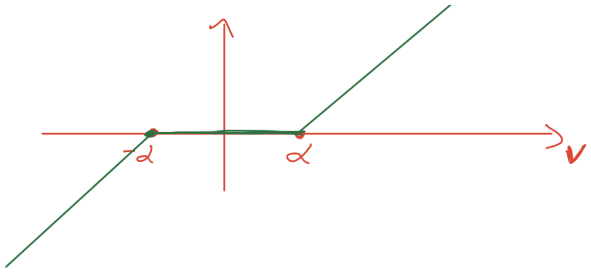

Figure 2: Graphical illustration of $S_\alpha(v)$

2 Linear and quadratic program

The standard form of a quadratic program (QP) is

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Px + q^\top x$$
$$\text{subject to } Ax = b,\ x \geq 0.$$

Here $x \geq 0$ means $x_i \geq 0$ for $i = 1, ..., n$. Also we assume that P is a PSD matrix. Note that when $P = 0$, we have a linear program (LP). Now we write the above problem in the ADMM form:

$$\min_{x,z} f(x) + g(z) \text{ subject to } Ax + Bz = c.$$

We can take $f(x)$ to be $\frac{1}{2}x^\top Px + q^\top x$ with domain being $\text{domain}(f) = \{x \mid Ax = b\}$, $g(z)$ to be $\mathbb{1}_{\mathbb{R}^n_+}(z)$, A, B to be I and $-I$, and c to be 0.

By playing with augmented Lagrangian, we can derive the scaled form of ADMM:

- $x^{(t+1)} = \text{argmin}_x(f(x) + \frac{\rho}{2}\|x - z^{(t)} + u^{(t)}\|^2)$.
- $z^{(t+1)} = \Pi_{\mathbb{R}^n_+}(x^{(t+1)} + u^{(t)})$.
- $u^{(t+1)} = u^{(t)} + x^{(t+1)} - z^{(t+1)}$.

In step one, we need to find $x^{(t+1)} = \text{argmin}_x(f(x) + \frac{\rho}{2}\|x - z^{(t)} + u^{(t)}\|^2)$. This optimization problem can be solved by solving the KKT system:

$$\begin{pmatrix} P + \rho I & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x^{(t+1)} \\ v \end{pmatrix} + \begin{pmatrix} q - \rho(z^{(t)} - u^{(t)}) \\ -b \end{pmatrix} = 0, \tag{1}$$

where v is the dual variable for this problem.

3 Last thoughts on ADMM

ADMM can also be used to solve problem where the cost function is not differentiable.

For example, while in least squares fitting, one tries to find x to minimize $\|Ax - b\|^2$. There is a variant (used in statistics, signal processing,) where we want to find x to minimize $\|Ax - b\|_1$. Here $\|z\|_1 = \sum_{i=1]}^n |z_i|$. In ADMM form, we can write

$$\min_{x,z} \|z\|_1 \text{ subject to } Ax - z = b.$$

So an ADMM algorithm iin scaled form for solving this is

- $x^{(t+1)} = (A^\top A)^{-1} A^\top (b + z^{(t)} - u^t)$.
- $z^{(t+1)} = \text{argmin}_z L_\rho(x^{(t+1)}, z, u^t)$.
- $u^{(t+1)} = u^{(t)} + Ax^{(t+1)} - b$.

The second step $z^{(t+1)} = \text{argmin}_z L_\rho(x^{(t+1)}, z, u^t)$ can be solved by "soft thresholding"

$$S_\alpha = \begin{cases} v - \alpha, & v > \alpha \\ 0, & |v| \leq \alpha \\ \alpha + v & v < -\alpha. \end{cases}$$

The graph of $S_\alpha(v)$ is illustrated in Figure 2. In practice,

$$z^{(t+1)} = S_{1/\rho}(Ax^{(t+1)} - b + u^{(t)}).$$

Math 173B - Lecture 19: Course Recap

1 Course Recap

The overall goal of this course was to introduce algorithms to minimize functions, where the algorithms were "light" in the sense that they could be scaled to problems with high dimension and lots of data.

1.1 Basics of this course

- Convexity/Strong Convexity
- Taylor
- Probability
- GD

1.2 SGD

The goal here is to minimize a function $F(x)$ of the form $\frac{1}{N}\sum_{i=1}^{N} f_i(x)$. The general update rule is

$$x^{(t+1)} = x^{(t)} - \alpha_t g^{(t)},$$

where $g^{(t)}$ has to be an unbiased estimate of $\nabla F(x^{(t)})$, i.e. $\mathbb{E}g^{(t)} = \nabla F(x^{(t)})$.

1.2.1 Vanilla SGD

In the vanilla form of SGD, the iteration reads

$$x^{(t+1)} = x^{(t)} - \alpha_t \nabla f_{i_t}(x^{(t)}),$$

where i_t is a randomly selected index.

For vanilla SGD, we have several convergence guarantees.

- With a fixed step size α, it will "converge" to a ball around the solution.
- With decreasing step size α_t, we can derive better convergence guarantees.

Then we compared SGD to GD in terms of

- Complexity.
- Convergence.

1.2.2 Variations on SGD

We introduced mini-batching SGD. The iteration reads

$$x^{(t+1)} = x^{(t)} - \alpha_t \frac{1}{m} \sum_{k=1}^{m} \nabla f_{i_k}(x^{(t)}),$$

where indices $i_1, ..., i_m$ are randomly sampled from $\{1, ..., N\}$. One can notice when $m = 1$, mini-batching SGD degenerates to vanilla SGD.

1.3 Optimization under constraints

The goal is to solve

$$\begin{gathered} \min f(x) \text{ subject to} \\ g_i(x) \leq 0, \text{ for } i = 1, ..., m, \\ h_i(x) = 0 \text{ for } i = 1, ..., p. \end{gathered}$$

1.3.1 Duality Theory

Here are some important concepts you should know:

- The Lagrangian $L(x, \lambda, \nu)$.
- Lagrangian dual function $f(\lambda, \nu) = \min_x L(x, \lambda, \nu)$.
- Primal and Dual optimization problem. Importantly, the dual problem is

$$\max_{\lambda, \nu} f(\lambda, \nu) \text{ subject to } \lambda \geq 0.$$

- Duality theory, which includes
  - weak duality
  - strong duality
  - Slater's condition
  - Complementary slackness
  - KKT conditions

Then we introduced algorithms for solving constrained optimization problems.

1.3.2 Dual Ascent

Suppose we want to solve

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

The algorithm reads

- $x^{(t+1)} = \text{argmin}_x L(x, y^{(t)})$.
- $y^{(t+1)} = y^{(t)} + \alpha_t (Ax^{(t+1)} - b)$.

The advantage of this algorithm is that when $f(x) = \sum_{i=1}^{n} f_i(x_i)$, you can write for $i = 1, ..., n$ in parallel:

- $x_i^{(t+1)} = \text{argmin}_{x_i} L_i(x_i, y^{(t)})$.
- $y^{(t+1)} = y^{(t)} + \alpha_t (Ax^{(t+1)} - b)$.

1.3.3 Augmented Lagrangian

We can define the augmented Lagrangian:

$$L_\rho(x, y) = f(x) + y^\top (Ax - b) + \frac{\rho}{2}\|Ax - b\|^2.$$

1.3.4 Method of Multipliers

The iteration reads:

- $x^{(t+1)} = \text{argmin}_x L_\rho(x, y^{(t)})$.
- $y^{(t+1)} = y^{(t)} + \rho(Ax^{(t+1)} - b)$.

1.3.5 ADMM

Here the goal is to solve a speical class of constrained optimization problem:

$$\min_{x,z} f(x) + g(z) \text{ subject to } Ax + Bz = c.$$

The iteration reads:

- $x^{(t+1)} = \text{argmin}_x L_\rho(x, z^{(t)}, y^{(t)})$.
- $z^{(t+1)} = \text{argmin}_z L_\rho(x^{(t+1)}, z, y^{(t)})$.
- $y^{(t+1)} = y^{(t)} + \rho(Ax^{(t+1)} + Bz^{(t+1)} - c)$.

Here are the things you should know about ADMM:

- The scaled form of ADMM.
- Convergence results.
- Examples of problems where ADMM is used:
  - Proximal operator
  - general convex constraints
  - projection onto convex sets
  - quadratic programs