■ 0.1 Interior Point Method for Linear Programs

In this section, we study interior point methods. In practice, this method reduces the problem of optimizing a convex function to solving a small number (often less than 30) linear systems. In theory, it reduces the problem to $\tilde{O}(\sqrt{n})$ linear systems.

We start by describing interior point method for linear programs. We establish a polynomial bound and then later discuss implementation details.

■ 0.1.1 Basic Properties

We first consider the primal problem

$$(\mathrm{P}): \qquad \min_x c^\top x \text{ subject to } Ax = b, x \geq 0$$

where $A \in^{m \times n}$. The difficulty of linear programs is the constraint $x \geq 0$. Without this constraint, we can simply solve it as a linear system. One natural idea to solve linear programs is to replace the hard constraint $x \geq 0$ by some smooth function. So, let us consider the following "regularized" version of the linear program for some $t \geq 0$:

$$(\mathrm{P}_t): \qquad \min_x c^\top x - t \sum_{i=1}^n \ln x_i \text{ subject to } Ax = b.$$

We will explain the reason of choosing $\ln x$ in more detail later. For now, we can think it as a nice function that blows up to ∞ as x approaches zero.

One can think that $-\ln x$ gives a force from every constraint $x \geq 0$ to make sure $x \geq 0$ is true. Since the gradient of $-\ln x$ blows up when $x = 0$, when x is close enough, the force is large enough to counter the cost c. When $t \to 0$, then the problem (P_t) is closer to the original problem (P) and hence the minimizer of (P_t) is closer to a minimizer of (P).

First, we give a formula for the minimizer of (P_t).

Lemma 0.1 (Existence and Uniqueness of central path)**.** *If the polytope $\{Ax = b, x \geq 0\}$ has an interior, then the optimum of (P_t) is unique and is given by the solution of the following system:*

$$\begin{aligned} xs &= t, \\ Ax &= b, \\ A^\top y + s &= c, \\ (x, s) &\geq 0 \end{aligned}$$

where the variables s_i are additional "slack" variables, and $xs = t$ is shorthand of $x_i s_i = t$ for all i.

Proof. The optimality condition, using dual variables y for the Lagrangian of $Ax = b$ is given by

$$c - \frac{t}{x} = A^\top y.$$

Write $s_i = \frac{t}{x_i}$, to get the formula. The solution is unique because the function $-\ln x$ is strictly convex. □

Definition 0.2. We define the central path $\mathcal{C}_t = (x^{(t)}, y^{(t)}, s^{(t)})$ as the sequence of points satisfying

$$\begin{aligned} x^{(t)} s^{(t)} &= t, \\ Ax^{(t)} &= b, \\ A^\top y^{(t)} + s^{(t)} &= c, \\ (x^{(t)}, s^{(t)}) &\geq 0. \end{aligned}$$

To give another interpretation of the central path, note that the dual problem is

$$(\mathrm{D}): \qquad \max_{y,s} b^\top y \text{ subject to } A^\top y + s = c, s \geq 0.$$

Note that for any feasible x and y, we have that

$$0 \leq c^\top x - b^\top y = c^\top x - x^\top A^\top y = x^\top s.$$

Hence, (x, y, s) solves the linear program if it satisfies the central path equation with $t = 0$. Therefore, following the central path is a balanced way to decrease $x_i s_i$ uniformly to 0. We can formalize the intuition that for small t, $x^{(t)}$ is a good approximation of the primal solution. In fact t itself is a bound on the error of the current solution.

Lemma 0.3 (Duality Gap)**.** *We have that*

$$\textit{Duality Gap} = c^\top x^{(t)} - b^\top y^{(t)} = c^\top x^{(t)} - \left(x^{(t)}\right)^\top A^\top y^{(t)} = \left(x^{(t)}\right)^\top s^{(t)} = tn.$$

The interior point method follows the following framework:

1. Find $\mathcal{C}_1$
2. Until $t < \frac{\varepsilon}{n}$,
   (a) Use $\mathcal{C}_t$ to find $\mathcal{C}_{(1-h)t}$ for $h = \frac{1}{10\sqrt{n}}$.

Note that this algorithm only finds a solution with ε error. If the linear program is integral, we can simply stop at small enough ε and round it off to the closest integral point.

■ 0.1.2 Following the central path

During the algorithm, we maintain a point (x, y, s) such that $Ax = b$, $A^\top y + s = c$ and $x_i s_i$ is close to t for all i. We show how to find a feasible $(x + \delta_x, y + \delta_y, s + \delta_s)$ such that it is even closer to t. We can write the equation as follows:

$$\begin{aligned} (x + \delta_x)(s + \delta_s) &\approx t, \\ A(x + \delta_x) &= b, \\ A^\top(y + \delta_y) + (s + \delta_s) &= c. \end{aligned}$$

(Omitted the non-negative conditions.) Using our assumption on (x, y, s) and noting that $\delta_x \cdot \delta_s$ is small, the equation can simplified as follows. We use the notation $X = (x)$, $S = (s)$.

$$\begin{bmatrix} 0 & A^\top & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \delta_x \\ \delta_y \\ \delta_s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ t - xs \end{bmatrix}.$$

This is a linear system and hence we can solve it exactly.

Exercise 0.4. Let $r = t - xs$. Prove that $S\delta_x = (I - \overline{P})r$ and $X\delta_s = \overline{P}r$ where $\overline{P} = XA^\top(AS^{-1}XA^\top)^{-1}AS^{-1}$.

First, we show that $x = x + \delta_x$ and $s = s + \delta_s$ are feasible.

Lemma 0.5. *Suppose* $\sum_i (x_i s_i - t)^2 \leq \varepsilon^2 t^2$ *with* $\varepsilon < \frac{1}{2}$*. Then,*$x_i > 0$ *and* $s_i > 0$ *for all* i*.*

Proof. Note that $P^2 = P$. However, in general $P \neq P^\top$,i.e. P might not be an orthogonal projection matrix. It will be convienient to consider the orthogonal projection matrix $P = S^{-\frac{1}{2}}X^{\frac{1}{2}}A^\top(AS^{-1}XA^\top)^{-1}AS^{-\frac{1}{2}}X^{\frac{1}{2}}$. Note that

$$X^{-1}\delta_x = S^{-\frac{1}{2}}X^{-\frac{1}{2}}(I - P)S^{-\frac{1}{2}}X^{-\frac{1}{2}}r.$$

By the assumption for each i, $x_i s_i \geq (1-\varepsilon)t$. Therefore, we have

$$\begin{aligned}\|X^{-1}\delta_x\|_2 &\leq \frac{1}{\sqrt{(1-\epsilon)t}}\|(I-P)S^{-\frac{1}{2}}X^{-\frac{1}{2}}r\|_2\\ &\leq \frac{1}{\sqrt{(1-\epsilon)t}}\|S^{-\frac{1}{2}}X^{-\frac{1}{2}}r\|_2\\ &\leq \frac{1}{(1-\epsilon)t}\|r\|_2 \leq \frac{\epsilon}{1-\epsilon}.\end{aligned}$$

Similarly, we have $S^{-1}\delta_s = S^{-\frac{1}{2}}X^{-\frac{1}{2}}PS^{-\frac{1}{2}}X^{-\frac{1}{2}}r$. Hence, we have $\|S^{-1}\delta_s\|_2 \leq \frac{\epsilon}{1-\epsilon}$. Therefore, when $\epsilon < \frac{1}{2}$, we have both $\|X^{-1}\delta_x\|_\infty$ and $\|S^{-1}\delta_s\|_\infty$ less than 1, which shows that both x and s are positive.

Next, we show that xs is closer to t after one Newton step. □

Lemma 0.6. *If $\sum_i (x_i s_i - t)^2 \leq \varepsilon^2 t^2$ with $\epsilon < \frac{1}{4}$, we have that*

$$\sum_i (x_i s_i - t)^2 \leq \left(\epsilon^4 + 16\epsilon^5\right) t^2.$$

Proof. We have that $x_i\delta_{s,i} + s_i\delta_{x,i} = t - x_i s_i$. Using this,

$$\text{LHS} = \sum_i (x_i s_i - t)^2 = \sum_i (x_i s_i + x_i\delta_{s,i} + s_i\delta_{x,i} + \delta_{x,i}\delta_{s,i} - t)^2 = \sum_i \delta_{x,i}^2\delta_{s,i}^2 \leq ((1+\epsilon)t)^2 \cdot \sum_i \left(\frac{\delta_{x,i}}{x_i}\right)^2\left(\frac{\delta_{s,i}}{s_i}\right)^2$$

where in the last step we used $x_i^2 s_i^2 \leq (1+\varepsilon)^2 t^2$. Using the previous lemma, we have that

$$\begin{aligned}\text{LHS} &\leq ((1+\epsilon)t)^2 \cdot \|X^{-1}\delta_x\|_4^2\|S^{-1}\delta_s\|_4^2\\ &\leq ((1+\epsilon)t)^2 \cdot \|X^{-1}\delta_x\|_2^2\|S^{-1}\delta_s\|_2^2\\ &\leq ((1+\epsilon)t)^2\left(\frac{\epsilon}{1-\epsilon}\right)^4\\ &\leq \left(\epsilon^4 + 16\epsilon^5\right)t^2\end{aligned}$$

□

Using this, we have the main theorem.

Theorem 0.7. *We can solve a linear program to within δ "error" (see Lemma 0.8) in $O(\sqrt{n}\log(\frac{1}{\delta}))$ iterations and each iteration only needs to solve a linear system.*

Proof. Let $\Phi = \sum_i (x_i s_i - t)^2$ be the error of the current iteration. We always maintain $\Phi \leq \frac{t^2}{16}$ for the current (x, y, s) and t. At each step, we use Lemma 0.6 which makes $\Phi \leq \frac{t^2}{50}$. Then, we decrease t to $t(1 - \frac{1}{10\sqrt{n}})$. Note that

$$\sum_i (x_i s_i - t(1-h))^2 \leq 2\Phi + 2t^2h^2n \leq \frac{2t^2}{50} + \frac{2t^2}{100} \leq \frac{(t(1-h))^2}{16}.$$

Therefore, the invariant is preserved after each step. Since t is decreased by a $(1 - \frac{1}{10\sqrt{n}})$ factor each step, it takes $O(\sqrt{n}\log(\frac{1}{\delta}))$ to decrease t from 1 to δ^2. □

■ 0.1.3 Finding the initial point

The first question is to find $\mathcal{C}_1$. This can be handled by extending the problem to slightly higher dimension. To the reader familiar with the Simplex method, this might be reminiscent of the two phases of the simplex method, where the purpose of the first phase is to find a feasible initial solution.

Lemma 0.8. *Consider a linear program $\min_{Ax=b,x\geq 0} c^\top x$ with n variables and d constraints. Assume that*

1. *Diameter: For any $x \geq 0$ with $Ax = b$, we have that $\|x\|_\infty \leq R$.*
2. *Lipschitz constant of the objective: $\|c\|_\infty \leq L$.*

For any $0 < \delta \leq 1$, the modified linear program $\min_{\overline{A}\overline{x}=\overline{b},\overline{x}\geq 0} \overline{c}^\top \overline{x}$ with

$$\overline{A} = \begin{bmatrix} A & 0 & \frac{1}{R}b - A1_n \\ 1_n^\top & 1 & 0 \end{bmatrix}, \overline{b} = \begin{bmatrix} \frac{1}{R}b \\ n+1 \end{bmatrix}, \text{ and } \overline{c} = \begin{bmatrix} \delta/L \cdot c \\ 0 \\ 1 \end{bmatrix}$$

satisfies the following:

1. $\overline{x} = \begin{bmatrix} 1_n \\ 1 \\ 1 \end{bmatrix}, \overline{y} = \begin{bmatrix} 0_d \\ -1 \end{bmatrix}$ *and* $\overline{s} = \begin{bmatrix} 1_n + \frac{\delta}{L} \cdot c \\ 1 \\ 1 \end{bmatrix}$ *are feasible primal dual vectors.*
2. *For any feasible primal dual vectors $(\overline{x}, \overline{y}, \overline{s})$ with duality gap at most δ^2, the vector $\hat{x} = R \cdot \overline{x}_{1:n}$ ($\overline{x}_{1:n}$ are the first n coordinates of $\overline{x}$) is an approximate solution to the original linear program in the following sense*

$$\begin{aligned} c^\top \hat{x} &\leq \min_{Ax=b,x\geq 0} c^\top x + LR \cdot \delta, \\ \|A\hat{x} - b\|_1 &\leq 4n\delta \cdot \left(R \sum_{i,j} |A_{i,j}| + \|b\|_1 \right), \\ \hat{x} &\geq 0. \end{aligned}$$

Part 1. For the first result, straightforward calculations show that $(\overline{x}, \overline{y}, \overline{s}) \in^{(n+2)\times(d+1)\times(n+2)}$ are feasible, i.e.,

$$\overline{A}\overline{x} = \begin{bmatrix} A & 0 & \frac{1}{R}b - A1_n \\ 1_n^\top & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1_n \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{R}b \\ n+1 \end{bmatrix} = \overline{b}$$

and

$$\begin{aligned} \overline{A}^\top \overline{y} + \overline{s} &= \begin{bmatrix} A^\top & 1_n \\ 0 & 1 \\ \frac{1}{R}b^\top - 1_n^\top A^\top & 0 \end{bmatrix} \cdot \begin{bmatrix} 0_d \\ -1 \end{bmatrix} + \begin{bmatrix} 1_n + \frac{\delta}{L} \cdot c \\ 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} -1_n \\ -1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1_n + \frac{\delta}{L} \cdot c \\ 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{\delta}{L} \cdot c \\ 0 \\ 1 \end{bmatrix} = \overline{c} \end{aligned}$$

Part 2. For the second result, we let

$$\text{OPT} = \min_{Ax=b,x\geq 0} c^\top x, \quad \text{and,} \quad \overline{\text{OPT}} = \min_{\overline{A}\overline{x}=\overline{b},\overline{x}\geq 0} \overline{c}^\top \overline{x}$$

For any optimal $x \in^n$ in the original LP, we consider the following $\overline{x} \in^{n+2}$

$$\overline{x} = \begin{bmatrix} \frac{1}{R}x \\ n+1-\frac{1}{R}\sum_{i=1}^n x_i \\ 0 \end{bmatrix} \tag{1}$$

and $\overline{c} \in^{n+2}$

$$\overline{c} = \begin{bmatrix} \frac{\delta}{L} \cdot c^\top \\ 0 \\ 1 \end{bmatrix} \tag{2}$$

We want to argue that $\overline{x} \in^{n+2}$ is feasible in the modified LP. It is obvious that $\overline{x} \geq 0$, it remains to show $\overline{A}\overline{x} = \overline{b} \in^{d+1}$. We have

$$\overline{A}\overline{x} = \begin{bmatrix} A & 0 & \frac{1}{R}b - A1_n \\ 1_n^\top & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{R}x \\ n+1-\frac{1}{R}\sum_{i=1}^n x_i \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{R}Ax \\ n+1 \end{bmatrix} = \begin{bmatrix} \frac{1}{R}b \\ n+1 \end{bmatrix} = \overline{b},$$

where the third step follows from $Ax = b$, and the last step follows from definition of $\overline{b}$.

Therefore, using the definition of $\overline{x}$ in (1) we have that

$$\overline{\mathrm{OPT}} \leq \overline{c}^\top \overline{x} = \begin{bmatrix} \frac{\delta}{L} \cdot c^\top & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{R}x \\ n+1-\frac{1}{R}\sum_{i=1}^n x_i \\ 0 \end{bmatrix} = \frac{\delta}{LR} \cdot c^\top x = \frac{\delta}{LR} \cdot \mathrm{OPT}. \tag{3}$$

where the first step follows from modified program is solving a minimization problem, the second step follows from definition of $\overline{x} \in^{n+2}$ (1) and $\overline{c} \in^{n+2}$ (2), the last step follows from $x \in^n$ is an optimal solution in the original linear program.

Given a feasible $(\overline{x}, \overline{y}, \overline{s}) \in^{(n+2)\times(d+1)\times(n+2)}$ with duality gap δ^2, we can write $\overline{x} = \begin{bmatrix} \overline{x}_{1:n} \\ \tau \\ \theta \end{bmatrix} \in^{n+2}$ for some $\tau \geq 0$, $\theta \geq 0$. We can compute $\overline{c}^\top \overline{x}$ which is $\frac{\delta}{L} \cdot c^\top \overline{x}_{1:n} + \theta$. Then, we have

$$\frac{\delta}{L} \cdot c^\top \overline{x}_{1:n} + \theta \leq \overline{\mathrm{OPT}} + \delta^2 \leq \frac{\delta}{LR} \cdot \overline{\mathrm{OPT}} + \delta^2, \tag{4}$$

where the first step follows from definition of duality gap, the last step follows from (3).

Hence, we can upper bound the $\overline{\mathrm{OPT}}$ of the transformed program as follows:

$$c^\top \hat{x} = R \cdot c^\top \overline{x}_{1:n} = \frac{LR}{\delta} \cdot \frac{\delta}{L} c^\top \overline{x}_{1:n} \leq \frac{RL}{\delta}(\frac{\delta}{LR} \cdot \mathrm{OPT} + \delta^2) = \mathrm{OPT} + LR \cdot \delta,$$

where the first step follows by $\hat{x} = R \cdot \overline{x}_{1:n}$, the third step follows by (4).

Note that

$$\frac{\delta}{L} c^\top \overline{x}_{1:n} \geq -\frac{\delta}{L}\|c\|_\infty \|\overline{x}_{1:n}\|_1 = -\frac{\delta}{L}\|c\|_\infty \|\frac{1}{R}x\|_1 \geq -\frac{\delta}{L}\|c\|_\infty \frac{n}{R}\|x\|_\infty \geq -\delta n, \tag{5}$$

where the second step follows from definition $\overline{x} \in^{n+2}$, and the last step follows from $\|c\|_\infty \leq L$ and $\|x\|_\infty \leq R$.

We can upper bound the θ in the following sense,

$$\theta \leq \frac{\delta}{LR} \cdot \overline{\mathrm{OPT}} + \delta^2 + \delta n \leq 2n\delta + \delta^2 \leq 4n\delta \tag{6}$$

where the first step follows from (4) and (5), the second step follows by $\mathrm{OPT} = \min_{Ax=b, x\geq 0} c^\top x \leq nLR$ (because $\|c\|_\infty \leq L$ and $\|x\|_\infty \leq R$), and the last step follows from $\delta \leq 1 \leq n$.

The constraint in the new polytope shows that

$$A\overline{x}_{1:n} + (\frac{1}{R}b - A1_n)\theta = \frac{1}{R}b.$$

Using $\hat{x} = Rx_{1:n} \in^n$, we have

$$A\frac{1}{R}\hat{x} + (\frac{1}{R}b - A1_n)\theta = \frac{1}{R}b.$$

Rewriting it, we have $A\hat{x} - b = (RA1_n - b)\theta \in^d$ and hence

$$\|A\hat{x} - b\|_1 = \|(RA1_n - b)\theta\|_1 \leq \theta(\|RA1_n\|_1 + \|b\|_1) \leq \theta \cdot (R\|A\|_1 + \|b\|_1) \leq 4n\delta \cdot (R\|A\|_1 + \|b\|_1),$$

where the second step follows from triangle inequality, the third step follows from $\|A1_n\|_1 \leq \|A\|_1$ (because the definition of entry-wise ℓ_1 norm), and the last step follows from (6).

0.1.4 Why $\sqrt{n}$?

The central path is the solution to the following ODE

$$\begin{aligned} S_t \frac{d}{dt} x_t + X_t \frac{d}{dt} s_t &= 1, \\ A \frac{d}{dt} x_t &= 0, \\ A^\top \frac{d}{dt} y_t + \frac{d}{dt} s_t &= 0. \end{aligned}$$

Solving this linear system, we have that $S_t \frac{dx_t}{dt} = (I - P_t)1$ and $X_t \frac{ds_t}{dt} = P_t 1$ where $P_t = X_t A^\top (A S_t^{-1} X_t A^\top)^{-1} A S_t^{-1}$. Using that $x_t s_t = t$, we have that

$$P_t = X_t A^\top (A X_t^2 A^\top)^{-1} A X_t = S_t^{-1} A^\top (A S_t^{-2} A^\top)^{-1} A S_t^{-1}$$

and that $X_t^{-1} \frac{dx_t}{dt} = \frac{1}{t}(I - P_t)1$ and $S_t^{-1} \frac{ds_t}{dt} = \frac{1}{t} P_t 1$. Equivalently, we have

$$\frac{d \ln x_t}{d \ln t} = (I - P_t)1 \text{ and } \frac{d \ln s_t}{d \ln t} = P_t 1.$$

Note that

$$P_t 1_\infty \leq P_t 1_2 = \sqrt{n}.$$

Hence, x_t and s_t can change by at most a constant factor when we change t by a $1 \pm \frac{1}{\sqrt{n}}$ factor.

Exercise 0.9. If we are given x such that $\ln x - \ln x_{t\infty} = O(1)$, then we can find x_t by solving $\tilde{O}(1)$ linear systems.

0.2 Interior Point Method for Convex Programs

The interior point method can be used to optimize any convex function. For more in-depth treatment, please see the structural programming section in [?].

Recall that any convex optimization problem

$$\min_x f(x)$$

can be rewritten in the epigraph form as

$$\min_{\{(x,t):\ f(x) \leq t\}} t.$$

Hence, it suffices to study the problem $\min_{x \in K} c^\top x$. Similar to the case of linear programs, we replace the hard constraint $x \in K$ by a soft constraint as follows:

$$\min_x \phi_t(x) \text{ where } \phi_t(x) = t c^\top x + \phi(x).$$

where $\phi(x)$ is a convex function such that $\phi(x) \to +\infty$ as $x \to \partial K$. Note that we put the parameter t in front of the cost $c^\top x$ instead of ϕ as in the last lecture, it is slightly more convenient here. We say ϕ is a *barrier* for K. To be concrete, we can always keep in mind $\phi(x) = -\sum_{i=1}^n \ln x_i$. As before, we define the central path.

Definition 0.10. The central path $x_t = \arg\min_x \phi_t(x)$.

The interior point method follows the following framework:

1. Find x close to x_1.
2. While t is not tiny,
   (a) Move x closer to x_t
   (b) $t \to (1 + h) \cdot t$.

0.2.1 Newton method and self-concordance

In this section, we give a general analysis for the Newton method. In the next section, we will use this to show that interior point method can be generalized to convex optimization. A key property of the Newton method is that it is invariant under linear transformation. In general, whenever a method uses k^{th} order information, we need to assume the k^{th} derivative is continuous. Otherwise, the k^{th} derivative is not useful for algorithmic purposes. For the Newton method, it is convenient to assume that the Hessian is Lipschitz. Since the method is invariant under linear transformation, it only makes sense to impose an assumption that is invariant under linear transformation.

Definition 0.11. Given a convex function $f :^n \rightarrow$, and any point $x \in^n$, define the norm $._x$ as

$$v_x^2 = v^\top \nabla^2 f(x) v.$$

We call a function f self-concordant if for any $h \in$ and any x in f, we have

$$D^3 f(x)[h, h, h] \leq 2h_x^3$$

where $D^k f(x)[h_1, h_2, \cdots, h_k]$ is the directional k^{th} derivative of f along the directions $h_1, h_2, \cdots, h_k$.

Remark. The constant 2 is chosen so that $-\ln(x)$ exactly satisfies the assumption and it is not very important, in that by scaling f, we can change any constant to any other constant.

Exercise 0.12. Show that the following property is equivalent fo self-concordance as defined above: restricted on any straight line $g(t) = f(x + th)$, we have $g'''(t) \leq 2g''(t)^{3/2}$.

Exercise 0.13. Show that the functions $x^\top Ax$, $-\ln x$, $-\ln(1 - \sum x_i^2)$, $-\ln \det X$ are self-concordant under suitable nonnegativity conditions.

The self-concordance condition says that locally, the Hessian does not change too fast, i.e., the change in the Hessian is bounded by its magnitude (to the power 1.5). We will skip the proof of the lemma below.

Lemma 0.14. *Given a self-concordant function f, for any $h_1, h_2, h_3 \in$, we have*

$$D^3 f(x)[h_1, h_2, h_3] \leq 2h_{1x} h_{2x} h_{3x}.$$

From the self-concordance condition, we have the following more directly usable property.

Lemma 0.15. *For a self-concordant function f and any $x \in f$ and any $y - x_x < 1$, we have that*

$$(1 - y - x_x)^2 \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \frac{1}{(1 - y - x_x)^2} \nabla^2 f(x).$$

Proof. Let $\alpha(t) = \left\langle \nabla^2 f(x + t(y - x))u, u \right\rangle$. Then, we have that

$$\alpha'(t) = D^3 f(x + t(y - x))[y - x, u, u].$$

By self-concordance, we have

$$|\alpha'(t)| \leq 2y - x_{x+t(y-x)} u_{x+t(y-x)}^2. \tag{7}$$

For $u = y - x$, we have $|\alpha'(t)| \leq 2\alpha(t)^{\frac{3}{2}}$. Hence, we have $\frac{d}{dt} \frac{1}{\sqrt{\alpha(t)}} \geq -1$. Integrating both sides wrt t, we have

$$\frac{1}{\sqrt{\alpha(t)}} \geq \frac{1}{\sqrt{\alpha(0)}} - t = \frac{1}{x - y_x} - t.$$

Rearranging it gives

$$y - x_{x+t(y-x)}^2 = \alpha(t) \leq \frac{1}{(\frac{1}{x-y_x} - t)^2} = \frac{x - y_x^2}{(1 - tx - y_x)^2}.$$

For general u, (7) gives

$$|\alpha'(t)| \leq 2\frac{x-y_x}{1-tx-y_x}\alpha(t).$$

Rearranging,

$$\left|\frac{d}{dt}\ln\alpha(t)\right| \leq 2\frac{x-y_x}{1-tx-y_x} = -2\frac{d}{dt}\ln(1-tx-y_x)$$

Integrating both from $t=0$ to 1 gives the result. □

Now we are ready to study the convergence of Newton method:

Lemma 0.16. *Given a self-concordant convex function f, consider the iteration*

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

Suppose that $r = \nabla f(x)_{\nabla^2 f(x)^{-1}} < 1$, then we have

$$\nabla f(x')_{\nabla^2 f(x')^{-1}} \leq \frac{r^2}{(1-r)^2}.$$

Remark 0.17. Note that $\nabla f(x)_{\nabla^2 f(x)^{-1}} = \nabla^2 f(x)^{-1}\nabla f(x)_x$ is the step size of the Newton method. This is a measurement of the error, since the goal is to find x with $\nabla f(x) = 0$.

Proof. Lemma 0.15 shows that

$$\nabla^2 f(x') \succeq (1-r)^2\nabla^2 f(x).$$

and hence

$$\nabla f(x')_{\nabla^2 f(x')^{-1}} \leq \frac{\nabla f(x')_{\nabla^2 f(x)^{-1}}}{(1-r)}.$$

To bound $\nabla f(x')$, we calculate that

$$\begin{aligned}\nabla f(x') &= \nabla f(x) + \int_0^1 \nabla^2 f(x+t(x'-x))(x'-x)dt\\ &= \nabla f(x) - \int_0^1 \nabla^2 f(x+t(x'-x))(\nabla^2 f(x))^{-1}\nabla f(x)dt\\ &= \left(\nabla^2 f(x) - \int_0^1 \nabla^2 f(x+t(x'-x))dt\right)(\nabla^2 f(x))^{-1}\nabla f(x). \qquad (8)\end{aligned}$$

For the term in the bracket, we use Lemma 0.15 to get (note that $\int_0^1(1-tr)^2dt = 1-r+\frac{1}{3}r^2$):

$$(1-r+\frac{1}{3}r^2)\nabla^2 f(x) \preceq \int_0^1 \nabla^2 f(x+t(x'-x))dt \preceq \frac{1}{1-r}\nabla^2 f(x).$$

Therefore, we have

$$\left\|(\nabla^2 f(x))^{-\frac{1}{2}}\left(\nabla^2 f(x) - \int_0^1 \nabla^2 f(x+t(x'-x))dt\right)(\nabla^2 f(x))^{-\frac{1}{2}}\right\|_{\text{op}} \leq \max(\frac{r}{1-r}, r-\frac{1}{3}r^2) = \frac{r}{1-r}.$$

Putting it into (8) gives

$$\begin{aligned}\nabla f(x')_{\nabla^2 f(x)^{-1}} &= \|\nabla^2 f(x)^{-\frac{1}{2}}\nabla f(x')\|_2\\ &\leq \left\|(\nabla^2 f(x))^{-\frac{1}{2}}\left(\nabla^2 f(x) - \int_0^1 \nabla^2 f(x+t(x'-x))dt\right)(\nabla^2 f(x))^{-\frac{1}{2}}\right\|_{\text{op}}\|(\nabla^2 f(x))^{-\frac{1}{2}}\nabla f(x)\|_2\\ &\leq \frac{r}{1-r}\cdot r\\ &= \frac{r^2}{1-r}.\end{aligned}$$

□

Finally, we bound the error of the current iterate in terms of $\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}}$.

Lemma 0.18. *Given x such that $\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} \leq \frac{1}{6}$, we have that*

- $\|x - x^*\|_{x^*} \leq 2\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}}$,
- $\|x - x^*\|_{x} \leq \frac{4}{3}\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}}$,
- $f(x) \leq f(x^*) + \|\nabla f(x)\|^2_{\nabla^2 f(x)^{-1}}$.

Proof. Let $r = \|x - x^*\|_x$. Suppose that $r \leq \frac{1}{4}$. Note that

$$\nabla f(x) = \nabla f(x) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + t(x - x^*))(x - x^*) dt.$$

Using that $\nabla^2 f(x^* + t(x - x^*)) \succeq (1 - (1 - t)r)^2 \nabla^2 f(x)$ (Lemma 0.15), we have

$$\begin{aligned}
\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} &= \int_0^1 \nabla^2 f(x^* + t(x - x^*))(x - x^*)\, dt_{\,\nabla^2 f(x)^{-1}} \\
&\geq \int_0^1 (1 - (1 - t)r)^2 \|x - x^*\|_x dt \\
&= \left(1 - r + \frac{r^2}{3}\right) r \geq \frac{3r}{4}.
\end{aligned}$$

Using $\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} \leq \frac{1}{6}$, we have indeed $r \leq \frac{1}{4}$ (our lower bound above is a non-decreasing function) Using Lemma 0.15 again, we have

$$\|x - x^*\|_{x^*} \leq \frac{\|x - x^*\|_x}{1 - r} \leq \frac{4}{3}\frac{1}{1 - \frac{1}{4}}\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} \leq 2\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}}.$$

For the bound for $f(x)$, we have that

$$f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \int_0^1 (1 - t)(x - x^*)^\top \nabla^2 f(x^* + t(x - x^*))(x - x^*) dt.$$

Using that $\nabla^2 f(x^* + t(x - x^*)) \preceq \frac{1}{(1-(1-t)r)^2}\nabla^2 f(x)$, we have

$$\begin{aligned}
f(x) &\leq f(x^*) + \int_0^1 \frac{1 - t}{(1 - (1 - t)r)^2} dt \cdot \|x - x^*\|_x^2 \\
&= f(x^*) + \frac{1}{r^2}(\frac{r}{1 - r} + \log(1 - r))\|x - x^*\|_x^2 \\
&\leq f(x^*) + (\frac{1}{2} + r)\|x - x^*\|_x^2 \\
&\leq f(x^*) + \nabla f(x)^2_{\nabla^2 f(x)^{-1}}
\end{aligned}$$

where we used $r \leq \frac{1}{4}$ at the end. □

■ 0.2.2 Self-concordant barrier functions

To analyze the algorithm above, we need to assume that ϕ is well-behaved. We measure the quality of ϕ by $\|\nabla \phi\|_{\nabla^2 \phi(x)^{-1}}$. One can think this as the Lipschitz constant of ϕ but measured in the local norm.

Definition 0.19. We call ϕ a ν-self-concordant barrier for K if ϕ is self-concordant, $\phi(x) \to +\infty$ as $x \to \partial K$ and that $\|\nabla \phi(x)\|^2_{\nabla^2 \phi(x)^{-1}} \leq \nu$ for all x.

Not all convex functions are self-concordant. However, for our purpose, it suffices to show that we can construct a self-concordant barrier for any convex set.

Theorem 0.20. *Any convex set has an n-self concordant barrier.*

Unfortunately, this is an existence result and the barrier function is expensive to compute.In practice, we construct self-concordant barriers out of simpler ones:

Lemma 0.21. *We have the following self-concordant barriers. We use ν-sc as a short form for ν-self-concordant barrier.*

- $-\ln x$ *is* 1*-sc for* $\{x \geq 0\}$.
- $-\ln \cos(x)$ *is* 1*-sc for* $\{|x| \leq \frac{\pi}{2}\}$.
- $-\ln(t^2 - x^2)$ *is* 2*-sc for* $\{t \geq x_2\}$.
- $-\ln \det X$ *is* n*-sc for* $\{X \in^{n\times n}, X \succeq 0\}$.
- $-\ln x - \ln(\ln x + t)$ *is* 2*-sc for* $\{x \geq 0, t \geq -\ln x\}$.
- $-\ln t - \ln(\ln t - x)$ *is* 2*-sc for* $\{t \geq e^x\}$.
- $-\ln x - \ln(t - x \ln x)$ *is* 2*-sc for* $\{x \geq 0, t \geq x \ln x\}$.
- $-2\ln t - \ln(t^{2/p} - x^2)$ *is* 4*-sc for* $\{t \geq |x|^p\}$ *for* $p \geq 1$.
- $-\ln x - \ln(t^p - x)$ *is* 2*-sc for* $\{t^p \geq x \geq 0\}$ *for* $0 < p \leq 1$.
- $-\ln t - \ln(x - t^{-1/p})$ *is* 2*-sc for* $\{x > 0, t \geq x^{-p}\}$ *for* $p \geq 1$.
- $-\ln x - \ln(t - x^{-p})$ *is* 2*-sc for* $\{x > 0, t \geq x^{-p}\}$ *for* $p \geq 1$.

The following lemma shows how we can combine barriers.

Lemma 0.22. *If ϕ_1 and ϕ_2 are ν_1 and ν_2-self concordant barriers for K_1 and K_2 respectively, then $\phi_1 + \phi_2$ is a $\nu_1 + \nu_2$ self concordant barrier for $K_1 \cap K_2$.*

Lemma 0.23. *If ϕ is a ν-self concordant barrier for K, then $\phi(Ax+b)$ is ν-self concordant for $\{y : Ay+b \in K\}$.*

Exercise 0.24. Using the lemmas above, prove that $-\sum_{i=1}^m \ln(a_i^\top x - b_i)$ is an m-self concordant barrier for the convex set $\{Ax \geq b\}$.

■ 0.2.3 Main Algorithm and Analysis

Algorithm 1: `InteriorPointMethod`

```
Input: A $\nu$-self-concordant barrier $\phi$ for $K$, the minimizer $x$ of $\phi$.
Define $f_t(x) = tc^\top x + \phi(x)$. $t = \frac{1}{6}\|c\|^{-1}_{\nabla^2\phi(x)^{-1}}$.
while $t \leq \frac{\nu+\sqrt{\nu}}{\epsilon}$ do
    $x \leftarrow x - \nabla^2 f_t(x)^{-1}\nabla f_t(x)$.
    $t \leftarrow (1+h)t$ with $h = \frac{1}{9\sqrt{\nu}}$.
end
return $x$.
```

We first explain the termination condition. Intuitively, we can see that $\min f_t(x_t)$ tends to optimality as $t \to \infty$. We first need a lemma showing that the gradient of ϕ is small.

Lemma 0.25 (Duality Gap). *Suppose that ϕ is a ν-self concordant barrier. For any $x, y \in K$, we have that*

$$\langle \nabla\phi(x), y - x\rangle \leq \nu.$$

Proof. Let $\alpha(t) = \langle \nabla\phi(z_t), y - x\rangle$ where $z_t = x + t(y-x)$. Then, we have

$$\alpha'(t) = \left\langle \nabla^2\phi(z_t)(y-x), y-x \right\rangle .$$

Note that

$$\alpha(t) \leq \nabla\phi(z_t)_{\nabla^2\phi(z_t)^{-1}} y - x_{\nabla^2\phi(z_t)} \leq \sqrt{v} y - x_{\nabla^2\phi(z_t)}.$$

Hence, we have $\alpha'(t) \geq \frac{1}{v}\alpha(t)^2$. If $\alpha(0) \leq 0$, then we are done. Otherwise, α is increasing and hence $\alpha(1) > 0$. Since $\frac{1}{\alpha(1)} \leq \frac{1}{\alpha(0)} - \frac{1}{v}$. So, $\alpha(0) \leq v$. □

Lemma 0.26 (Duality Gap). *Suppose that ϕ is a ν-self concordant barrier, we have that*

$$\langle c, x_t\rangle \leq \langle c, x^*\rangle + \frac{\nu}{t}.$$

More generally, for any x such that $\|tc + \nabla\phi(x)\|_{(\nabla^2\phi(x))^{-1}} \leq \frac{1}{6}$, we have that

$$\langle c, x\rangle \leq \langle c, x^*\rangle + \frac{\nu + \sqrt{\nu}}{t}.$$

Proof. Let x^* be a minimizer of $c^\top x$ on K. By optimality, we have $tc + \nabla\phi(x) = 0$. Therefore, we have

$$\langle c, x_t\rangle - \langle c, x^*\rangle = \frac{1}{t}\left\langle \nabla\phi(x_t), x^* - x_t\right\rangle \leq \frac{\nu}{t}.$$

For the second result, let $f(x) = tc^T x + \phi(x)$. Lemma 0.18 shows that

$$\|x - x_t\|_x \leq 2\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} \leq \frac{1}{3}.$$

Hence,

$$\langle c, x - x_t\rangle \leq \|c\|_{\nabla^2\phi(x)^{-1}}\|x - x_t\|_x \leq \frac{1}{3}\|c\|_{\nabla^2\phi(x)^{-1}}$$

Using $c = \frac{tc + \nabla\phi(x)}{t} - \frac{\nabla\phi(x)}{t}$, we have

$$\begin{aligned}\langle c, x - x_t\rangle &\leq \frac{1}{3t}\left(\|tc + \nabla\phi(x)\|_{\nabla^2\phi(x)^{-1}} + \|\nabla\phi(x)\|_{\nabla^2\phi(x)^{-1}}\right)\\ &\leq \frac{1}{3t}\left(\frac{1}{6} + \sqrt{\nu}\right) \leq \frac{\sqrt{\nu}}{t}.\end{aligned}$$

This gives the result. □

Hence, it suffices to end with $t = (\nu + \sqrt{\nu})/\varepsilon$, which is exactly same as the previous lecture.

Theorem 0.27. *Given a ν-self concordant barrier ϕ and its minimizer. We can find $x \in K$ such that $c^\top x \leq c^\top x^* + \epsilon$ in*

$$O(\sqrt{\nu}\log(\frac{\nu}{\epsilon}\|c\|_{\nabla^2\phi(x)^{-1}}))$$

iterations.

Proof. We prove by induction that $\|\nabla f_t(x)\|_{\nabla^2 f_t(x))^{-1}} \leq \frac{1}{6}$ at the beginning of each iteration. This is true at the beginning by the definition of initial t. By Lemma 0.16, after the Newton step, we have

$$\|\nabla f_t(x)\|_{\nabla^2 f_t(x)^{-1}} \leq (\frac{1/6}{1 - 1/6})^2 = \frac{1}{25}.$$

Let $t' = (1+h)t$ with $h = \frac{1}{9\sqrt{\nu}}$. Note that $\nabla f_{t'}(x) = (1+h)tc + \nabla\phi(x)$ and hence $\nabla f_{t'}(x) = (1+h)\nabla f_t(x) - h\nabla\phi(x)$ and $\nabla^2 f_{t'}(x) = \nabla^2 f_t(x) = \nabla^2\phi(x)$. Therefore, we have that

$$\begin{aligned}\|\nabla f_{t'}(x)\|_{\nabla^2 f_{t'}(x)^{-1}} &= \|(1+h)\nabla f_t(x) - h\nabla\phi(x)\|_{\nabla^2 f_t(x)^{-1}}\\ &\leq (1+h)\|\nabla f_t(x)\|_{\nabla^2 f_t(x)^{-1}} + h\|\nabla\phi(x)\|_{\nabla^2\phi(x)^{-1}}\\ &\leq \frac{1+h}{25} + h\sqrt{\nu} \leq \frac{1}{6}.\end{aligned}$$

This completes the induction. □