Lecture 18

18.1 Second-Order Optimization

In the final part of the course, we study *second-order optimization* algorithms. These methods utilize the Hessian of the objective (or its approximation) to better capture the geometry of the problem. While second-order algorithms are typically more computationally expensive, they often possess superior convergence rates compared to first-order ones. Therefore, selecting an appropriate optimization algorithm requires a trade-off between iteration complexity and per-iteration cost; this choice depends on problem properties (such as dimension, sparsity, etc.) and the desired target accuracy.

We will be solving an optimization problem of the form,

$$\min_{x \in Q} f(x), \tag{18.1}$$

and distinguish between two principal cases, for each developing corresponding second-order algorithms:

1. *Unconstrained optimization.* $Q = \mathbb{R}^n$, where $\mathbb{R}^n$ is the target vector space. Note that in this setting, the vector space can be readily replaced by an *affine space*, $Q = \{x \in \mathbb{R}^n \ : \ Ax = b\}$, which is useful if we need to cover the affine equality constraints.

   First-order methods. We allow to compute $f(x)$ and $\nabla f(x)$ and perform very simple operations, such as *summation of two vectors*:

   $$x^+ \quad = \quad x - \alpha \nabla f(x). \tag{18.2}$$

   Or, choosing a distance function d (which has to be *simple*), the mirror descent step:

   $$\nabla d(x^+) \quad = \quad \nabla d(x) - \alpha \nabla f(x).$$

   These operations can be generalized further by the framework of composite optimization, e.g. treating an additive composite regularizer $\psi(y)$ to the objective in (18.1) by steps:

   $$x^+ \quad = \quad \arg\min_y \Big[\ \langle \nabla f(x), y - x \rangle + \beta_d(x; y) + \psi(y) \ \Big],$$

   where $\beta_d(x; y)$ is the Bregman divergence. However, all operations remain simple and we typically hope to obtain an explicit formula for x^+.

Second-order methods. The key assumption of second-order algorithms is

$$\textit{We can solve linear systems: } Hx = g.$$

Indeed, thanks to advances in linear algebra and the rapid development of numerical packages, we can use efficient linear algebra techniques (LU, QR, Cholesky, tridiagonal decomposition, SVD, ...). Note that efficient packages, such as LAPACK (integrated in Python and other languages), are able to solve linear systems of dimension $n \approx 2 \cdot 10^4$ in under a minute on a standard laptop (see Fig. 18.1). Along with time complexity, the main bottleneck for larger n becomes memory constraints (a dense $n \times n$ matrix of `float64` values with $n = 28000$ requires about 6 GB of RAM only to store it).

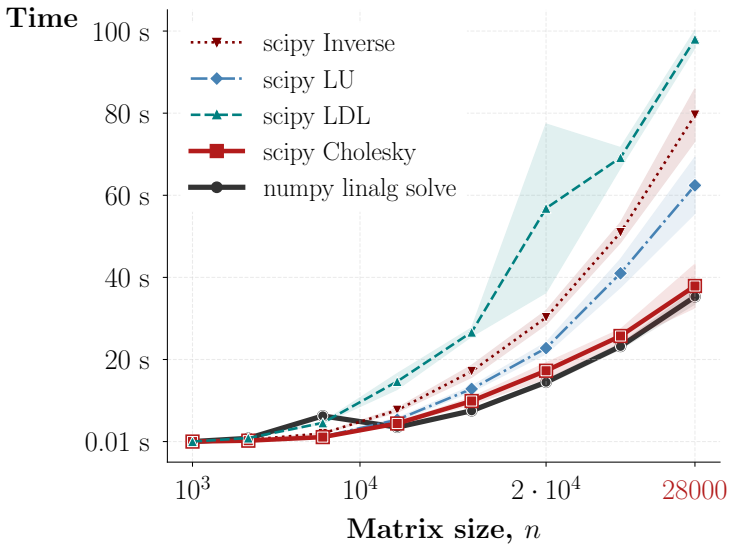

Figure 18.1: *Scaling of linear system solvers.* Solving a symmetric positive definite dense linear system in Python on a MacBook Pro 2024, using the following functions:
1. Inverting the matrix with `scipy.linalg.inv`
2. General system solver (LU decomposition) with `scipy.linalg.solve(assume_a='gen')`
3. Symmetric system solver (LDL decomposition) with `scipy.linalg.solve(assume_a='sym')`
4. Cholesky decomposition with `scipy.linalg.solve(assume_a='pos')`
5. Numpy default solver with `np.linalg.solve`.

An alternative, highly effective approach to solving a symmetric positive-definite linear system is to run a first-order method on a quadratic objective (e.g., the conjugate gradient method, or, in the case of constraints or non-smooth regularizers, the fast gradient method) — this scales well to very large values of n, as the first-order method requires only a procedure to compute Hessian-vector products, and we can perform only a few iterations of the solver to obtain an approximate solution. Further performance gains can be achieved if the linear system is *sparse* or *low-rank*.

Thus, the main idea behind second-order algorithms is to rely more on efficient linear algebra rather than solely on the summation of vectors, as in first-order methods. We will also assume access to a second-order local oracle, computing $f(x)$, $\nabla f(x)$, and the Hessian $\nabla^2 f(x)$.

We will study a few variants of Newton's method that achieve *provably better global rates* than those of the first-order methods, for convex and non-convex unconstrained optimization.

2. *Structured constrained optimization.* In this case, $Q \subset \mathbb{R}^n$ is a complicated convex set with a specific structure, while the objective function is linear, $f(x) = \langle c, x \rangle$.

   Our main assumption is the possibility of constructing a *self-concordant barrier* F for the set Q (see Fig. 18.2), which we study in detail in the following lectures. It appears that such barriers can be constructed for practically all known classes of convex optimization problems (e.g., linear programming, quadratic programming, semidefinite programming, etc.), and, moreover, they can be minimized very efficiently with second-order methods, achieving polynomial-time complexity and excellent practical performance.

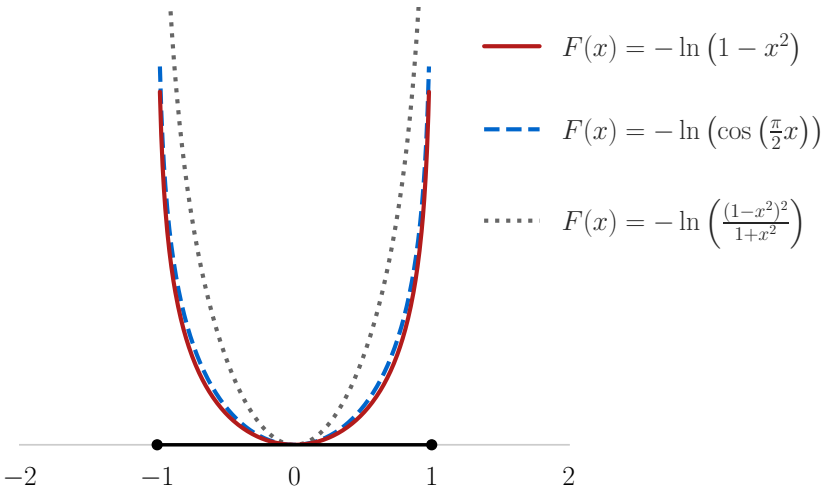

Figure 18.2: Self-concordant barriers for the segment $[-1, 1]$. See Lecture 21 for the definition.

This class of algorithms is known as *interior-point methods*, which places second-order algorithms in a unique position as an essential tool for the minimization of self-concordant barriers, and for the overall success of this framework.

18.1.1 Quadratic Taylor Approximation: Newton's Step

The idea of the classical Newton method is to use second-order (quadratic) Taylor expansion of the objective $f(y)$, around the current point $x \in \mathbb{R}^n$:

$$f(y) \quad = \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + o(\|y - x\|^2).$$

Then, we choose the next point x^+ as a minimum of the second-order model:

$$x^+ \quad = \quad \arg\min_{y \in \mathbb{R}^n} \Big[\, \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle \, \Big]. \tag{18.3}$$

Note that the minimum (18.3) might not exist at all, or if exists, it might not be unique.

For now, and for the next few lectures we might assume that $\nabla^2 f(x) \succ 0$, and then x^+ exists and unique as the minimizer of strongly convex model in (18.3).

The optimality condition for (18.3) states that x^+ should satisfy the linear equation:

$$\nabla f(x) + \nabla^2 f(x)(x^+ - x) \quad = \quad 0.$$

or, rearranging the terms, we obtain the classic Newton's step:

$$x^+ \;=\; x - \nabla^2 f(x)^{-1} \nabla f(x). \tag{18.4}$$

18.1.2 Affine Invariance

Let $x := Ay + b$ for an invertible matrix A and define new function

$$F(y) \;=\; f(Ay + b) \;=\; f(x).$$

Consider Newton's step (18.4) for the original objective, and Newton's step for the new function, from some point $y \in \mathbb{R}^n$:

$$y^+ \;=\; y - \nabla^2 F(y)^{-1} \nabla F(y)$$

It appears that Newton's step is *affine-invariant*:

Proposition 18.1.1. *Let $x = Ay + b$. Then, $x^+ = Ay^+ + b$.*

Proof. Note that

$$\begin{aligned} \nabla F(y) &= A^\top \nabla f(Ay + b) = A^\top \nabla f(x), \\ \nabla^2 F(y) &= A^\top \nabla^2 f(Ay + b) A = A^\top \nabla^2 f(x) A. \end{aligned}$$

Then,

$$\begin{aligned} y^+ &= y - \nabla^2 F(y)^{-1} \nabla F(y) \\ &= y - A^{-1} \nabla^2 f(x)^{-1} A^{-\top} A^\top \nabla f(x) \\ &= y - A^{-1} \nabla^2 f(x)^{-1} \nabla f(x). \end{aligned}$$

Therefore,

$$Ay^+ + b \;=\; Ay + b - \nabla^2 f(x)^{-1} \nabla f(x) \;=\; x^+.$$

□

Hence, Newton's method is independent of the choice of coordinate system in $\mathbb{R}^n$. This property can also be seen directly from (18.3), as the value of Taylor polynomial does not depend on the actual choice of the inner product $\langle \cdot, \cdot \rangle$. If we change the coordinate system and run the method from the corresponding initial point, the result remains the same. In other words, classic Newton's method cannot be "accelerated" by finding a better coordinate system, unlike the gradient method.

Note however that the basic gradient step (18.2) is invariant to shifts and orthogonal transformations:

Proposition 18.1.2. *Let $x = Uy + b$, where $UU^\top = I$. Consider $F(y) = f(Uy + b)$ and two gradient steps $y^+ = y - \alpha \nabla F(y)$ and $x^+ = x - \alpha \nabla f(x)$. Then,*

$$x^+ \;=\; Uy^+ + b$$

Exercise 18.1.1. Prove this proposition. Show that the gradient method is not affine-invariant.

18.2 Definition of Self-Concordant Functions

The main result about Newton's method is its *local quadratic convergence*: when the point is sufficiently close to the optimum $x \approx x^\star$, the method doubles known digits of the solution with every step. We will prove this result using a modern affine-invariant analysis.

We consider differentiable function $f : Q \to \mathbb{R}$, where $Q \subseteq \mathbb{R}^n$ is an open convex set. We assume that $\nabla^2 f(x) \succ 0$ everywhere on Q, so f is strictly convex.

18.2.1 Local Euclidean Structure

The main component in the definition of self-concordant functions is the notion of the *local norm*. Note that at every point $x \in Q$, the Hessian of f defines the Euclidean structure on $\mathbb{R}^n$:

$$\langle u, v \rangle_x \;:=\; D^2 f(x)[u, v] \;=\; \langle \nabla^2 f(x) u, v \rangle, \tag{18.5}$$

where in the right hand side we use the standard inner product. In (18.5), the matrix $\nabla^2 f(x)$ depends on $\langle \cdot, \cdot \rangle$, while $D^2 f(x)$ and, consequently, $\langle \cdot, \cdot \rangle_x$ *do not depend on the coordinate system*.

Correspondingly, we can define the local norm, which is the norm generated by the Hessian of the objective:

$$\|h\|_x \;:=\; \langle h, h \rangle_x^{1/2} \;=\; \langle \nabla^2 f(x) h, h \rangle^{1/2}, \qquad x \in Q, h \in \mathbb{R}^n. \tag{18.6}$$

We will use this norm to characterize smoothness of the objective f.

18.2.2 Third Derivative

To understand the behavior of (18.6), we fix a direction $h \in \mathbb{R}^n$ and study the quadratic form

$$g(x) \;=\; \|h\|_x^2 \;=\; \langle \nabla^2 f(x) h, h \rangle \;>\; 0, \tag{18.7}$$

as a function of x. By continuity, it is clear that

$$\text{when} \quad y \;\approx\; x \quad \text{then} \quad \nabla^2 f(y) \;\approx\; \nabla^2 f(x). \tag{18.8}$$

Our goal is to provide a *quantitative* and *affine-invariant* characterization of (18.8). For an arbitrary perturbation $u \in \mathbb{R}^n$, we consider

$$\varphi(t) \;=\; \|h\|_{x+tu}^2 \;=\; \langle \nabla^2 f(x + tu) h, h \rangle,$$

for small $t > 0$. Note that φ is a scalar function. Its derivative at zero,

$$\varphi'(0) \;=\; D^3 f(x)[h, h, u] \;\in\; \mathbb{R},$$

shows how fast the local norm (18.7) changes at x. It is known that $D^3 f(x)$ is a trilinear symmetric form, as soon as f is sufficiently differentiable.

18.2.3 Self-Concordant Functions

We say that a function $f : Q \to \mathbb{R}$ is *self-concordant* with constant $M \geq 0$, if

$$D^3 f(x)[h, h, u] \;\leq\; M \|h\|_x^2 \|u\|_x, \qquad \forall h, u \in \mathbb{R}^n, x \in Q. \tag{18.9}$$

This inequality means that the third derivative is bounded by constant $M \geq 0$, but the "boundedness" is measured by the local norm at the same point x (which provides the *concordance*). As a result, the parameter M does not depend on the coordinate system.

If we subsitute $u := h$ (the same direction) into (18.9), we obtain the bound:

$$|D^3 f(x)[h,h,h]| \;\leq\; M\|h\|_x^3 \;=\; M\langle \nabla^2 f(x)h, h\rangle^{3/2}, \qquad \forall h \in \mathbb{R}^n, x \in Q. \tag{18.10}$$

However, it appears that the reverse implication holds as well! If the last inequality is satisfied, then (18.9) also holds. So, inequality (18.10) can be used as a definition of self-concordance. It is easier to check whether a function is self-concordant with the latter inequality. At the same time, our original definition (18.9) is better suitable for an analysis.

The equivalence of two definitions is a consequence of the following fact (see also Exercise 18.2.2).

Lemma 18.2.1. *Let $T[\cdot,\cdot,\cdot]$ be a trilinear symmetric form. Denote by $S = \{h : \langle h, h\rangle = 1\}$ the standard Euclidean unit sphere. Then,*

$$\max_{h,u \in S} T[h,h,u] \;=\; \max_{h \in S} T[h,h,h]. \tag{18.11}$$

Proof. Let $h^\star, u^\star \in S$ be any pair of maximizers (which clearly exists, as T is a continuous function and S is a compact set):

$$T^\star \;=\; \max_{h,u \in S} T[h,h,u] \;=\; T[h^\star, h^\star, u^\star]. \tag{18.12}$$

Denote $\theta = \langle h^\star, u^\star\rangle$. Without loss of generality we can assume $\theta \geq 0$. If $\theta = 1$ then (18.11) is proved. Hence, we assume that $0 \leq \theta < 1$.

Consider the linear form, $\langle \ell, u\rangle \equiv T[h^\star, h^\star, u]$. Thus, we have $\langle \ell, u\rangle \leq \langle \ell, u^\star\rangle = T^\star$, for all $u \in S$. By Cauchy-Schwartz inequality we conclude that $\ell = T^\star u^\star$. Thus, we get

$$T[h^\star, h^\star, h^\star] \;=\; \langle \ell, h^\star\rangle \;=\; \theta T^\star. \tag{18.13}$$

Similarly, consider the symmetric matrix A defined by the equation $\langle Ah, h\rangle \equiv T[u^\star, h, h]$, and we have $\langle Ah, h\rangle \leq \langle Ah^\star, h^\star\rangle = T^\star$, for all $h \in S$. Thus, by the spectral theorem, we conclude that $h^\star$ is the eigenvector of the matrix corresponding to the maximal eigenvalue: $Ah^\star = T^\star h^\star$, and, therefore,

$$T[u^\star, u^\star, h^\star] \;=\; \langle Ah^\star, u^\star\rangle \;=\; \theta T^\star. \tag{18.14}$$

Denote $v^\star := \frac{h^\star + u^\star}{\|h^\star + u^\star\|_2} \in S$ and note that $\|u^\star + h^\star\|_2^2 \;=\; 2(1+\theta)$. Then,

$$\begin{aligned} T[v^\star, v^\star, h^\star] \quad &= \quad \tfrac{1}{2(1+\theta)}\Bigl(T[h^\star,h^\star,h^\star] + T[u^\star,u^\star,h^\star] + 2T[h^\star,h^\star,u^\star]\Bigr) \\ &\overset{(18.12),(18.13),(18.14)}{=} \quad \tfrac{2\theta+2}{2(1+\theta)} T^\star \;=\; T^\star. \end{aligned} \tag{18.15}$$

Hence, the new triplet $(v^\star, v^\star, h^\star)$ preserves the optimal value of T, while *shrinking the distance*:

$$\tfrac{1}{2}\|v^\star - h^\star\|_2^2 \;=\; \Bigl(1 - \sqrt{\tfrac{1+\theta}{2}}\Bigr) \;\overset{(*)}{\leq}\; \Bigl(1 - \tfrac{1}{\sqrt{2}}\Bigr)\cdot(1-\theta) \;=\; \Bigl(1 - \tfrac{1}{\sqrt{2}}\Bigr)\tfrac{1}{2}\|h^\star - u^\star\|_2^2, \tag{18.16}$$

where $(*)$ follows from convexity of the function $\varphi(\theta) = 1 - \sqrt{\frac{1+\theta}{2}}$ for $0 \leq \theta \leq 1$.

Finally, consider the set of all maximizers of $T[h,h,u]$, which is a compact nonempty set:

$$\Omega \;=\; \Bigl\{(h,u) \in S \times S \;:\; T[h,h,u] \;=\; T^\star\Bigr\},$$

and a continuous function $\rho(h,u) \;=\; \|h - u\|_2$. Let $(h^\star, u^\star)$ be the minimizer of ρ over Ω. If $h^\star \neq u^\star$, by the previous reasoning, we can find a pair $(v^\star, h^\star)$ with a strictly smaller value of ρ, which contradicts that $(h^\star, u^\star)$ is the minimizer. Hence, $h^\star = u^\star$. $\square$

18.2.4 Examples and Basic Properties

First, let us start with the following three basic examples that show that the class of self-concordant functions is not empty and actually quite broad.

1. *Convex quadratic functions:* $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ is self-concordant with $\boxed{M = 0}$.

2. *Negative logarithm:* $f(x) = -\ln x$, for $x > 0$. Indeed,

$$f'(x) \;=\; -\tfrac{1}{x}, \quad f''(x) \;=\; \tfrac{1}{x^2}, \quad f'''(x) \;=\; -\tfrac{2}{x^3}.$$

   Hence,

$$|f'''(x)| \;=\; 2(f''(x))^{3/2},$$

   and we conclude that f is self-concordant with $\boxed{M = 2}$. Since logarithmic barriers play the key role in the theory of interior-point methods, self-concordant functions with $M = 2$ are often called *standard self-concordant*.

3. *Strongly convex functions with Lipschitz Hessian.* Let f have Lipschitz Hessian with constant $L > 0$ w.r.t a fixed norm, and let f be strongly convex with constant $\mu > 0$. Thus, for all $x, y \in Q$ and $h \in \mathbb{R}^n$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \;\le\; L\|x - y\| \qquad \text{and} \qquad \mu\|h\|^2 \;\le\; \|h\|_x^2.$$

   Then,

$$D^3 f(x)[h, h, h] \;\le\; L\|h\|^3 \;\le\; \tfrac{L}{\mu^{3/2}}\|h\|_x^3.$$

   Hence, we can take $M = \frac{L}{\mu^{3/2}}$. However, note that $L_{\|\cdot\|}$ and $\mu_{\|\cdot\|}$ depend on the choice of the norm $\|\cdot\|$, while M does not. Thus,

$$\boxed{M \;\le\; \inf_{\|\cdot\|}\Big(\tfrac{L_{\|\cdot\|}}{\mu_{\|\cdot\|}^{3/2}}\Big).}$$

Logarithmic barrier for semidefinite cone. The fact that $f(x) = -\ln x$, for $x > 0$, is self-concordant can be generalized to the cone of positive-definite symmetric matrices,

$$\mathbb{S}^n_+ \;=\; \Big\{ X \in \mathbb{R}^{n\times n} \;:\; X = X^\top \succeq 0 \Big\} \;\subset\; \mathbb{S}^n.$$

Example 18.2.2. Define, for $X \in \operatorname{int}\mathbb{S}^n_+$, the logarithmic barrier:

$$f(X) \;=\; -\ln\det X \;=\; -\sum_{i=1}^n \ln\lambda_i(X). \tag{18.17}$$

Then,

$$Df(X)[H] \;=\; -\operatorname{tr}(X^{-1}H) \qquad \Rightarrow \qquad \nabla f(X) \;=\; -X^{-1},$$

$$D^2 f(X)[H, H] \;=\; \operatorname{tr}(X^{-1}HX^{-1}H) \;=\; \operatorname{tr}(S^2), \qquad \text{where} \qquad S = X^{-1/2}HX^{-1/2} \;\in\; \mathbb{S}^n.$$

$$D^3 f(X)[H, H, H] \;=\; -2\operatorname{tr}(X^{-1}HX^{-1}HX^{-1}H) \;=\; -2\operatorname{tr}(S^3).$$

It remains to notice that

$$\operatorname{tr}(S^3) \;=\; \sum_{i=1}^n \lambda_i(S)^3 \;\le\; \sum_{i=1}^n |\lambda_i(S)|^3 \;\le\; \Big(\sum_{i=1}^n |\lambda_i(S)|^2\Big)^{3/2} \;=\; \Big(\operatorname{tr}(S^2)\Big)^{3/2},$$

where we used the inequality $\|\cdot\|_3 \le \|\cdot\|_2$. Hence, $f(X)$ is self-concordant with constant $\boxed{M = 2}$.

Summation of self-concordant functions. For a sum of m functions:

$$f(x) \;=\; \sum_{i=1}^{m} f_i(x),$$

where each f_i, $1 \le i \le m$ is self-concordant with constant $M_i \ge 0$, we have

$$\begin{aligned} D^3 f(x)[u]^3 \;&=\; \sum_{i=1}^{m} D^3 f_i(x)[u]^3 \;\overset{(18.10)}{\le}\; \sum_{i=1}^{m} M_i \Big(D^2 f_i(x)[u]^2 \Big)^{3/2} \\ &\le\; \max_{1\le i\le m} M_i \cdot \sum_{i=1}^{m} \Big(D^2 f_i(x)[u]^2 \Big)^{3/2} \;\le\; \max_{1\le i\le m} M_i \cdot \Big(\sum_{i=1}^{m} D^2 f_i(x)[u]^2 \Big)^{3/2} \\ &=\; \max_{1\le i\le m} M_i \cdot \Big(D^2 f(x)[u]^2 \Big)^{3/2}, \end{aligned}$$

where in the last inequality we used that $\|\cdot\|_{3/2} \le \|\cdot\|_1$. Thus, f is self-concordant with constant

$$\boxed{M \;=\; \max_{1\le i\le m} M_i.}$$

Example 18.2.3. The logarithmic barrier for $\mathbb{R}^n_{>0}$,

$$f(x) \;=\; -\sum_{i=1}^{n} \ln x^{(i)}, \tag{18.18}$$

is self-concordant with constant $\boxed{M = 2}$. Note that (18.18) can be viewed as a restriction of the logarithmic barrier (18.17) for the cone of positive definite matrices $\mathbb{S}^n_{\succ 0}$ onto the subset of diagonal matrices with positive entries, that is isomorphic to $\mathbb{R}^n_{>0}$.

Affine restrictions and affine substitutions do not affect self-concordance, as to show in the following exercise.

Exercise 18.2.1. Let $g(y) = f(Ay + b)$, where $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. Show that g is self-concordant with the same constant $\boxed{M_g = M_f}$.

Exercise 18.2.2. Using affine-invariance and Lemma 18.2.1, show that the definition of self-concordance along one direction (18.10) implies the definition along two arbitrary directions (18.9).

Exercise 18.2.3. Let $g(x) = cf(x)$, for $c > 0$. What will be the constant of self-concordance M_g for g? Show that for $M_f > 0$, we can always choose c such that $M_g = 2$, so any self-concordant function can be made "standard self-concordant" after an appropriate rescaling.

Example 18.2.4. The logarithmic barrier for the polyhedron $\{\langle a_1, x\rangle < b_1, \ldots, \langle a_m, x\rangle < b_m\}$:

$$f(x) = -\sum_{i=1}^{m} \ln(b_i - \langle a_i, x\rangle)$$

is self-concordant with constant $\boxed{M = 2}$.