Lecture 23

23.1 Functions with Lipschitz Hessian

The framework of self-concordant barriers and interior-point methods that we studied is very powerful, as it enables Newton's method to solve broad classes of convex structured problems with polynomial-time complexity.

The main drawback of this approach is that it requires a model designer to formulate the optimization problem in a specific form (linear programming, conic programming, etc.) In practice, we often encounter problems that are given in a *black-box* form, which may also be *non-convex*, and thus fail to satisfy the assumptions of the interior-point machinery.

Nevertheless, it is still possible to apply second-order methods in these cases by employing ideas from first-order optimization, thereby establishing superior convergence properties through the use of second-order information, i.e., the Hessians $\nabla^2 f(x)$.

23.1.1 Problem and Assumptions

We consider the minimization problem:

$$\min_{x \in \mathbb{R}^n} f(x), \tag{23.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function, possibly non-convex. The ideas that we will discuss can be generalized to the constrained or even the fully-composite case (Lecture 9). However, the most important and the simplest setting to study is unconstrained minimization (23.1).

We assume that the objective f is bounded from below:

$$f^\star \;\; := \;\; \inf_{x \in \mathbb{R}^n} f(x) > -\infty,$$

while our goal, in a general non-convex setting, is to find an *approximate stationary point* to f, as finding the global minimum is generally intractable from a complexity standpoint (Lecture 2).

We will see that by using the Hessians along with a stronger smoothness assumption on f, we can find a stationary point faster than with gradient descent. Moreover, we can additionally guarantee convergence to a *second-order stationary point* (i.e., points where the Hessian is nearly positive semidefinite), which helps to better distinguish between saddle points and local minima.

Let us fix a positive definite matrix $B = B^\top \succ 0$ (e.g., $B := I$, the identity matrix). We use it to define the *global norms* in our space, primal, dual and operator norms:

$$\|h\| \;\; := \;\; \langle Bh, h \rangle^{1/2} \;\; = \;\; \|B^{1/2} h\|_2, \qquad h \in \mathbb{R}^n,$$

$$\|g\|_* \;\; := \;\; \langle g, B^{-1} g \rangle^{1/2} \;\; = \;\; \|B^{-1/2} g\|_2, \qquad g \in \mathbb{R}^n,$$

and for a symmetric matrix $A = A^\top \in \mathbb{R}^{n \times n}$:

$$\begin{aligned} \|A\| &:= \max_{h \in \mathbb{R}^n \,:\, \|h\| \leq 1} \|Ah\|_* = \max_{h \in \mathbb{R}^n \,:\, \|h\| \leq 1} \langle Ah, h \rangle = \max_{u \in \mathbb{R}^n \,:\, \|u\|_2 \leq 1} \langle B^{-1/2} A B^{-1/2} u, u \rangle \\ &= \max\Big\{ \lambda_{\max}(B^{-1/2} A B^{-1/2}), -\lambda_{\min}(B^{-1/2} A B^{-1/2}) \Big\}. \end{aligned}$$

Lipschitz Hessian. Our main assumption is that f has a Lipschitz continuous Hessian, for some constant $L \geq 0$:

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq L\|y - x\|, \qquad x, y \in \mathbb{R}^n. \tag{23.2}$$

This condition is equivalent to

$$-L\|y - x\| B + \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \nabla^2 f(x) + L\|y - x\| B, \qquad x, y \in \mathbb{R}^n,$$

which can be seen as a variant of the Hessian stability (compare with Lemma 19.1.1 from Lecture 19 on self-concordant functions).

23.1.2 Taylor Approximation Bounds

Using assumption (23.2) we can characterize the global approximation error for the Taylor models, the linear model of the gradient:

$$\nabla f(y) \approx \nabla f(x) + \nabla^2 f(x)(y - x), \tag{23.3}$$

and the quadratic model of the function:

$$f(y) \approx Q(x; y) := f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle. \tag{23.4}$$

Note that approximations (23.3) and (23.4) are *local* in their nature: they serve as good models when $x \approx y$.

In contrast, assumption (23.2) is *global*, as it holds for all $x, y \in \mathbb{R}^n$. Integrating bound (23.2) we obtain the following:

Lemma 23.1.1. *It holds, for any $x, y \in \mathbb{R}^n$:*

$$\begin{aligned} \|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\|_* &\leq \tfrac{L}{2} \|y - x\|^2 \\ |f(y) - Q(x; y)| &\leq \tfrac{L}{6} \|y - x\|^3. \end{aligned}$$

Proof. By the main theorem of calculus, we have, for any h such that $\|h\| \leq 1$:

$$\begin{aligned} \langle \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x), h \rangle &= \int_0^1 \langle \nabla^2 f(x + \tau(y - x))(y - x) - \nabla^2 f(x)(y - x), h \rangle d\tau \\ &\leq \int_0^1 \|\nabla^2 f(x + \tau(y - x)) - \nabla^2 f(x)\| \cdot \|y - x\| d\tau \\ &\leq L\|y - x\|^2 \int_0^1 \tau d\tau = \tfrac{L}{2} \|y - x\|^2. \end{aligned}$$

And, by the integral form of the Taylor theorem, we have:

$$\begin{aligned} |f(y) - Q(x;y)| &= \int_0^1 (1-\tau)\langle \Big[\nabla^2 f(x+\tau(y-x)) - \nabla^2 f(x)\Big](y-x), y-x\rangle d\tau \\ &\leq L\|y-x\|^3 \int_0^1 (1-\tau)\tau d\tau = L\|y-x\|^3 \cdot (\tfrac{1}{2} - \tfrac{1}{3}) = \tfrac{L}{6}\|y-x\|^3. \end{aligned}$$

□

23.2 Cubic Regularization of Quadratic Model

From the previous lemma, we obtain the following *global upper model* of the objective function, around any point $x \in \mathbb{R}^n$:

$$\begin{aligned} f(y) &\leq \Omega_H(x;y) := Q(x;y) + \tfrac{H}{6}\|y-x\|^3 \\ &= f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle + \tfrac{H}{6}\|y-x\|^3, \end{aligned} \tag{23.5}$$

which holds *for all* $y \in \mathbb{R}^n$, when the regularization parameter $H \geq 0$ is sufficiently large. Indeed, by Lemma 23.1.1, inequality (23.5) holds uniformly for all $x, y \in \mathbb{R}^n$ at least when $H \geq L$.

Therefore, a natural idea for an optimization method is to minimize the upper model (23.5) to obtain the next iterate:

$$x^+ := \arg\min_{y \in \mathbb{R}^n} \Omega_H(x;y). \tag{23.6}$$

Note that $H := 0$ in (23.6) corresponds exactly to the *pure Newton step*. At the same time, when $H \geq L$, it follows from previous observations (23.5) that we can ensure *global progress* for each iterate; rearranging the terms, we have:

$$\begin{aligned} f(x) - f(x^+) &\overset{(23.5)}{\geq} -\Big[\langle \nabla f(x), x^+ - x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(x^+ - x), x^+ - x\rangle + \tfrac{H}{6}\|x^+ - x\|^3\Big] \\ &\overset{(23.6)}{=} \max_{y \in \mathbb{R}^n}\Big\{\langle \nabla f(x), x-y\rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle - \tfrac{H}{6}\|y-x\|^3\Big\}, \end{aligned}$$

and the last expression is strictly positive[1], unless $\nabla f(x) = 0$, i.e., we are already at a stationary point. In the latter case, we either remain at the same point ($x^+ = x$) if $\nabla^2 f(x) \succeq 0$, or we *jump out of it* if there exists a direction with a negative quadratic form value, $\nabla^2 f(x) \not\succeq 0$ (a strict saddle point or a strict local maximum).

Thus, iterations of the form (23.6) seem very attractive. However, notice that finding x^+ is not trivial, as the model $\Omega_H(x;y)$, as a function of y, is generally non-convex and might possess isolated local minima, as well as multiple global minima[2]. It is not a priori clear whether x^+ can be computed efficiently, which is the main question that we should ask to a new approach.

The following observations are important for making iterations (23.6) practical:

[1] Indeed, $y := x$ already results in a zero value, and a slight perturbation along the linear term will yield a positive value for the maximization subproblem.

[2] When $\nabla f(x) = 0$ and $\lambda_{\min}(\nabla^2 f(x)) < 0$, we can *jump out* of the stationary point x in multiple ways (see Exercise 23.2.1).

1. Notably, when the initial objective in (23.1) is *convex*, we know that $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^n$. Therefore, $\Omega_H(x; y)$ is a strictly convex function of y with a unique global minimizer x^+.

   Hence, we can apply any already known method from convex optimization to find x^+ (e.g., we can run the composite version of the fast gradient method, treating the cubic regularizer as the composite term, or the interior-point method, which requires building a suitable self-concordant barrier for the epigraph of the cubic model).

2. The optimality condition for (23.6) is the following non-linear equation[3]:

$$\nabla_y \Omega_H(x; x^+) \;=\; \nabla f(x) + \nabla^2 f(x)(x^+ - x) + \tfrac{H}{2} r B(x^+ - x) \;=\; 0, \tag{23.7}$$

   where $r := \|x^+ - x\|$. Thus, the global minimum x^+ must satisfy (23.7), though not every stationary point is the global minimum. In practice, using a stationary point that satisfies (23.7) in an algorithm might already be sufficient to ensure progress. Moreover, instead of an exact solution to (23.7), we can relax this condition to an *approximate stationary point*:

$$\nabla_y \Omega_H(x; x^+) \;\approx\; 0,$$

   and use, for example, the gradient descent algorithm to find such x^+.

3. Fortunately, it appears that we can *always compute* the global solution x^+ to (23.6) by using the structure of the subproblem and linear algebra techniques, such as the SVD of the Hessian. We discuss this approach further in more detail.

Exercise 23.2.1. Assume $\nabla f(x) = 0$ and $\lambda_{\min} := \lambda_{\min}(\nabla^2 f(x)) < 0$. Show that the set of all global minimizers of (23.6) consists of vectors

$$x^+ \;=\; x \pm \tau h,$$

where h is an eigenvector of the Hessian corresponding to the smallest eigenvalue: $\nabla^2 f(x) h = \lambda_{\min} h$, and $\tau > 0$ is a step-size that depends on $\lambda_{\min}$ and the regularization parameter $H > 0$.

23.2.1 Bound for the New Gradient

An immediate consequence of the optimality condition (23.7) is the following important lemma. Note that it holds even for $H := 0$ (the pure Newton step).

Lemma 23.2.1. *For any $H \geq 0$, it holds*

$$\|\nabla f(x^+)\|_* \;\leq\; \tfrac{L+H}{2} r^2. \tag{23.8}$$

Proof. Substituting the optimality condition into the bound on the gradient approximation in Lemma 23.1.1, we get

$$\|\nabla f(x^+) + \tfrac{H}{2} r B(x^+ - x)\|_* \overset{(23.7)}{=} \|\nabla f(x^+) - \nabla f(x) - \nabla^2 f(x)(x^+ - x)\|_* \leq \tfrac{L}{2} r^2.$$

Using triangle inequality gives (23.8). □

Inequality (23.8) allows us to compare the length of the step $r := \|x^+ - x\|$ with the norm of the gradient at the *new point*.

[3] It is nonlinear due to the presence of r, which depends on x^+.

Remark 23.2.2. Notice that when performing the gradient method step with a parameter $H > 0$:

$$\bar{x} \;=\; x - \tfrac{1}{H} B^{-1} \nabla f(x),$$

we have $\|\nabla f(x)\|_* = H\|\bar{x} - x\|$. Assuming that the gradient is Lipschitz with constant L_1, we obtain a similar bound to (23.8), but with a different power:

$$\|\nabla f(\bar{x})\|_* \;\leq\; \|\nabla f(x)\|_* + L_1\|\bar{x} - x\| \;=\; (L_1 + H)\|\bar{x} - x\|.$$

The difference in the power leads to different convergence rates.

23.2.2 Local Quadratic Convergence for Strongly Convex Functions

Before moving on to the general non-convex case, let us verify that the cubic regularization of the quadratic Taylor model will *preserve the local quadratic convergence* of the pure Newton method.

The local quadratic convergence is the most distinguishing feature of Newton's method, and we are definitely interested to keep it.

We assume that f is a strongly convex function, for a positive parameter $\mu > 0$:

$$\nabla^2 f(x) \;\succeq\; \mu B, \qquad x \in \mathbb{R}^n. \tag{23.9}$$

Let us multiply the optimality condition (23.7), which in this case defines the unique global minimum x^+, by $\langle \cdot, x^+ - x \rangle$. Rearranging the terms, we obtain

$$r\|\nabla f(x)\|_* \;\geq\; \langle \nabla f(x), x^+ - x \rangle \overset{(23.7)}{=} \langle \nabla f(x)(x^+ - x), x^+ - x \rangle + \tfrac{H}{2} r^3 \overset{(23.9)}{\geq} \mu r^2.$$

From this inequality we get the following bound.

Lemma 23.2.3. *For any $H \geq 0$, it holds:*

$$r \;\leq\; \tfrac{1}{\mu}\|\nabla f(x)\|_*. \tag{23.10}$$

It remains to combine (23.10) with (23.8):

$$\|\nabla f(x^+)\|_* \;\leq\; \tfrac{L+H}{2} r^2 \;\leq\; \tfrac{L+H}{2\mu^2}\|\nabla f(x)\|_*^2, \tag{23.11}$$

and this is the local quadratic convergence! As soon as the initial gradient is sufficiently small, the next gradient will be quadratically smaller, and we can estimate the region of the quadratic convergence as follows.

Theorem 23.2.4. *For any $H \geq 0$, the cubic Newton method converges quadratically in the local region:*

$$\mathcal{Q} \;:=\; \Big\{ x \;:\; \|\nabla f(x)\|_* \;\leq\; \tfrac{\mu^2}{L+H} \Big\}. \tag{23.12}$$

Thus, starting from $x_0 \in \mathcal{Q}$ and performing the iterations $x_{k+1} = \arg\min_{y \in \mathbb{R}^n} \Omega_H(x_k; y)$, $k \geq 0$, we get

$$\|\nabla f(x_k)\|_* \;\leq\; \varepsilon$$

after the following number of steps:

$$k \;=\; 1 + \Big\lceil \log_2 \log_2 \tfrac{2\mu^2}{(L+H)\varepsilon} \Big\rceil. \tag{23.13}$$

Proof. Indeed,

$$\frac{L+H}{2\mu^2}\|\nabla f(x_k)\|_* \overset{(23.11)}{\leq} \left(\frac{L+H}{2\mu^2}\|\nabla f(x_{k-1})\|_*\right)^2 \leq \ldots \leq \left(\frac{L+H}{2\mu^2}\|\nabla f(x_0)\|_*\right)^{2^k} \leq \left(\frac{1}{2}\right)^{2^k},$$

which gives (23.13). □

Note that this result holds for any $H \geq 0$, including $H := 0$. Therefore, we automatically reestablish the local quadratic convergence of the pure Newton method. Compared to the self-concordant analysis (see Lecture 20), the result of Theorem 23.2.4 is no longer affine-invariant — as is the case with the cubic Newton method — since we fix the coordinate system through the operator B.

It is remarkable that the region (23.12) of quadratic convergence is of the same order as covered by the general self-concordant theory for Newton's method on strongly convex functions with Lipschitz Hessian (see Corollary 20.1.4 in Lecture 20).

Hence, we see that the cubic regularization of Newton's method "does not harm" the best local quadratic approximation provided by Taylor's polynomial $Q(x; y)$.

23.3 Non-Convex Quadratics and Strong Duality

Let us discuss how the cubic subproblem (23.6) can be solved globally, even in the non-convex case. To this end, we consider the following more general problem using simplified notation:

$$\min_{y\in\mathbb{R}^n}\Big\{ P(y) \;=\; \langle g, y\rangle + \tfrac{1}{2}\langle Ay, y\rangle + \varphi(\langle By, y\rangle) \Big\}, \tag{23.14}$$

where $g \in \mathbb{R}^n$ is an arbitrary vector and $A = A^\top \in \mathbb{R}^{n\times n}$ is an arbitrary symmetric matrix, not necessary positive semidefinite, representing correspondingly the gradient and the Hessian from the cubic model.

As before, we assume that $B = B^\top \succ 0$, and φ is a non-decreasing univariate convex function defined on $\mathbb{R}_{\geq 0}$ and representing the regularizer. The most interesting examples are as follows:

1. *Cubic regularization* is covered by the choice $\varphi(s) := \frac{H}{6}s^{3/2}$. Indeed, substituting it into (23.14) and using that $\|y\| := \langle By, y\rangle^{1/2}$ gives

$$P(y) \;=\; \langle g, y\rangle + \tfrac{1}{2}\langle Ay, y\rangle + \tfrac{H}{6}\|y\|^3. \tag{23.15}$$

2. *Trust-region approach.* For a given parameter $r \geq 0$ (the trust-region radius), we set

$$\varphi(s) \;:=\; \begin{cases} 0, & s \leq r \\ +\infty, & \text{otherwise.} \end{cases}$$

Substituting it into (23.14) we obtain the trust-region subproblem:

$$\min_{y\in\mathbb{R}^n\,:\,\|y\|\leq r}\Big\{ \langle g, y\rangle + \tfrac{1}{2}\langle Ay, y\rangle \Big\}. \tag{23.16}$$

Trust-region methods are another popular way to globalize Newton's steps, especially for non-convex problems. They work by restricting the Taylor quadratic polynomial to a ball around the current iterate, the region where we "trust" our second-order model. Cubic and

trust-region subproblems can be seen as equivalent, up to the choice of the parameters H and r.

We mainly focus on cubic regularization, as our assumption on the Lipschitz continuity of the Hessian (23.2) immediately leads to the natural choice of the regularization parameter $H := L$. Historically, trust-region methods have been used intensively in practice, with a wide variety of efficient solvers developed specifically for subproblem (23.16). Given the shared structure (23.14), these efficient solvers can be used for the cubic case as well, with an appropriate change of φ.

3. *Quartic regularization.* One can think of other choices of φ, such as $\varphi(s) = \frac{H}{24}s^2$, which leads to the quartic model:
$$P(y) \;=\; \langle g, y\rangle + \tfrac{1}{2}\langle Ay, y\rangle + \tfrac{H}{24}\|y\|^4.$$
Thus, the global minimum of this function can be found with the same effort as in the cubic case (23.15).

For $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{+\infty\}$, we seek to use the following adjoint representation:

$$\varphi(s) \;=\; \max_{\tau \geq 0} \Big[\, \tau s - \varphi^\star(\tau) \, \Big], \tag{23.17}$$

where $\varphi^\star(\tau)$ is a convex function. A fundamental fact from convex analysis is that the function $\varphi^\star$ defined by (23.17) can be found as the *convex conjugate* of φ:

$$\varphi^\star(\tau) \;=\; \max_{s \geq 0} \Big[\, \tau s - \varphi(s) \, \Big].$$

In other words, Fenchel-Moreau duality holds: $\varphi^{\star\star}(s) = \varphi(s)$, when φ is "sufficiently good".

It is easy to check that for the case of cubic regularization, we use the following conjugate pair:

$$\varphi(s) \;=\; \tfrac{H}{6}s^{3/2} \qquad \text{and} \qquad \varphi^\star(\tau) \;=\; \tfrac{2^4}{3H^2}\tau^3.$$

Then, we can write the primal dual pair of problems[4]:

$$\begin{aligned}
\min_{y \in \mathbb{R}^n} P(y) \;&=\; \min_{y \in \mathbb{R}^n} \Big\{ \, \langle g, y\rangle + \tfrac{1}{2}\langle Ay, y\rangle + \varphi\big(\langle By, y\rangle\big) \, \Big\} \\
&\overset{(23.17)}{=} \min_{y \in \mathbb{R}^n} \max_{\tau \geq 0} \Big\{ \, \langle g, y\rangle + \tfrac{1}{2}\langle Ay, y\rangle + \tfrac{\tau}{2}\langle By, y\rangle - \varphi^\star\big(\tfrac{\tau}{2}\big) \, \Big\} \qquad (23.18) \\
&\geq \max_{\tau \geq 0} \min_{y \in \mathbb{R}^n} \Big\{ \, \langle g, y\rangle + \tfrac{1}{2}\langle Ay, y\rangle + \tfrac{\tau}{2}\langle By, y\rangle - \varphi^\star\big(\tfrac{\tau}{2}\big) \, \Big\} \;=\; \max_{\tau \in \mathcal{W}} D(\tau),
\end{aligned}$$

where

$$D(\tau) \;:=\; \min_{y \in \mathbb{R}^n} \Big\{ \, \langle g, y\rangle + \tfrac{1}{2}\langle (A + \tau B)y, y\rangle \, \Big\} - \varphi^\star\big(\tfrac{\tau}{2}\big) \;=\; -\tfrac{1}{2}\langle g, (A + \tau B)^{-1} g\rangle - \varphi^\star\big(\tfrac{\tau}{2}\big),$$

is a concave univariate function defined on the (possibly open) ray:

$$\mathcal{W} \;:=\; \{\tau \geq 0 \;:\; A + \tau B \succ 0\} \;=\; \{\tau \geq 0 \;:\; \tau > -\lambda_{\min}(B^{-1/2} A B^{-1/2})\}.$$

[4] We could use inf and sup here instead of min and max to be a bit more precise.

In fact, one can show that the *strong duality* holds in this case:

$$\boxed{\min_{y \in \mathbb{R}^n} P(y) \;=\; \max_{\tau \in \mathcal{W}} D(\tau)} \tag{23.19}$$

This is remarkable, as the problem in the left-hand-side of (23.19) is non-convex, while in the right-hand-side we have a simple maximization of a univariate concave function.

It appears that the primal problem albeit non-convex in its current form, possesses *hidden convexity*. This can be viewed as a consequence of some fundamental facts about interactions of quadratic forms — one can show that the joint image of two quadratic forms

$$U \;:=\; \Big\{ [u_1, u_2]^\top \;=\; \Big[\langle g, y \rangle + \tfrac{1}{2}\langle Ay, y \rangle,\; \langle By, y \rangle \Big]^\top \;:\; y \in \mathbb{R}^n \Big\} \;\subseteq\; \mathbb{R}^2,$$

is a convex set in two-dimensional space. Therefore, the original non-convex primal problem can be rewritten as *convex minimization*:

$$\min_{y \in \mathbb{R}^n} P(y) \;=\; \min_{u \in U} \Big\{ u_1 + \varphi(u_2) \Big\},$$

while being an implicit formulation.

23.3.1 Solving Cubic Subproblem in Practice

Consider, for simplicity, $B := I$, using the standard Euclidean norm $\|\cdot\|_2$ in the regularization. The generalization to arbitrary $B \succ 0$ is straightforward. To compute the cubic Newton step (23.6):

$$x \;\mapsto\; x^+ \;=\; \arg\min_{y \in \mathbb{R}^n} \Omega_H(x; y),$$

we perform the following steps:

1. Simplify the quadratic part, by computing the *eigenvalue* (or *tridiagonal*) decomposition of the Hessian $\nabla^2 f(x)$:
$$\nabla^2 f(x) \;=\; U \Lambda U^\top,$$
where the matrix U is orthogonal: $UU^\top = I$, and the matrix Λ is diagonal (or tridiagonal). This can be done in $\mathcal{O}(n^3)$ arithmetic operations.

2. Then, we can solve the *univariate dual problem*
$$\max_{\tau \geq 0} \Big\{ -\tfrac{1}{2}\langle \nabla f(x), (\nabla^2 f(x) + \tau I)^{-1} \nabla f(x) \rangle - \tfrac{2}{3H^2}\tau^3 \;:\; \tau \;>\; -\lambda_{\min}(\nabla^2 f(x)) \Big\},$$
e.g., by finding the root of the equation $D'(\tau^\star) = 0$. This gives us the following non-linear equation to solve, using the reparametrization $\tau^\star \equiv \frac{H}{2} r^\star$:
$$h(r^\star) \;=\; \|s(r^\star)\|_2 - r^\star \;=\; 0, \tag{23.20}$$
where
$$s(r) \;:=\; \big(\nabla^2 f(x) + \tfrac{H}{2} r I\big)^{-1} \nabla f(x).$$
Note that $\|s(r)\|_2$ is a monotonically decreasing convex function that we want to intersect with the identity function in order to find $r^\star$. For solving (23.20), we can employ either binary search or univariate Newton's method, which will use $\tilde{\mathcal{O}}(n^2)$ arithmetic operations, hiding logarithmic factors.

3. After finding the root $r^\star$ of $h(\cdot)$, one step of the cubic Newton method can be written in the following explicit form:

$$\boxed{x^+ \;=\; x - \left(\nabla^2 f(x) + \tfrac{H}{2} r^\star I\right)^{-1} \nabla f(x),}$$

unless we are in the rare *degenerate* case: $\frac{H}{2} r^\star = -\lambda_{\min}(\nabla^2 f(x))$, which corresponds to the situation when the supremum of the dual problem is achieved at the boundary of the open ray $\mathcal{W}$. This situation should be handled separately.