Lecture 25

25.1 Quasi-Self-Concordant Functions

25.1.1 Motivational Example: Smoothness of Loss Functions

Consider the following canonical problem in a separable form (e.g., training a generalized linear models such as the logistic regression):

$$\min_{x \in \mathbb{R}^n} \Big[f(x) \;\; := \;\; \sum_{i=1}^m \ell(\langle a_i, x \rangle) \Big] \tag{25.1}$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is a differentiable convex loss function. To which problem class does this objective belong?

For that, we look at the derivatives. The first derivative linear form:

$$\langle \nabla f(x), h \rangle \;\; = \;\; \sum_{i=1}^m \ell'(\langle a_i, x \rangle) \langle a_i, h \rangle, \qquad x, h \in \mathbb{R}^n, \tag{25.2}$$

and the second derivative quadratic form:

$$\|h\|_x^2 \;\; \equiv \;\; \langle \nabla^2 f(x) h, h \rangle \;\; = \;\; \sum_{i=1}^m \ell''(\langle a_i, x \rangle) \langle a_i, h \rangle^2, \qquad x, h \in \mathbb{R}^n. \tag{25.3}$$

First-order methods. Assuming that the loss has a uniformly bounded second derivative, $\ell''(t) \leq L_1$, for all $t \in \mathbb{R}$, we can bound the Hessian of f as follows:

$$\langle \nabla^2 f(x) h, h \rangle \;\; \leq \;\; L_1 \sum_{i=1}^m \langle a_i, h \rangle^2 \;\; \equiv \;\; L_1 \langle Bh, h \rangle \;\; \equiv \;\; L_1 \|h\|_B^2 \;\; \leq \;\; L_1 \|B\| \cdot \|h\|_2^2, \tag{25.4}$$

where $B := \sum_{i=1}^m a_i a_i^\top$ is a symmetric positive semidefinite matrix. Without loss of generality, we can assume $B \succ 0$. Indeed, if for a certain direction $h \in \mathbb{R}^n$ it holds that $Bh = 0$, then both $\langle \nabla f(x), h \rangle = 0$ and $\langle \nabla^2 f(x) h, h \rangle = 0$ simultaneously, and a method will be automatically restricted to move only along the subspace spanned by $\{a_1, \ldots, a_m\}$.

Therefore, we conclude that f has the Lipschitz gradient, and we can apply first-order methods to (25.1). For example, the complexity of the basic gradient method to find a point x_k such that $f(x_k) - f^\star \leq \varepsilon$, starting from an arbitrary $x_0 \in \mathbb{R}^n$ is

$$O\Big(\tfrac{L_1 \|x_0 - x^\star\|_B^2}{\varepsilon}\Big) \;\; \leq \;\; O\Big(\tfrac{L_1 \|B\| \cdot \|x_0 - x^\star\|_2^2}{\varepsilon}\Big) \tag{25.5}$$

first-order oracle calls, correspondingly, when using the norm $\|\cdot\|_B$ or $\|\cdot\|_2$ in the method. These complexities can be improved by extracting the square root with the fast gradient method.

Second-order methods. What about second-order methods? For them, we need to bound the third derivative, for any $x, h, u \in \mathbb{R}^n$:

$$\begin{aligned} |D^3 f(x)[h,h,u]| &= \Big|\sum_{i=1}^m \ell'''(\langle a_i, x\rangle)\langle a_i, h\rangle^2 \cdot \langle a_i, u\rangle\Big| \\ &\le \max_{1\le i\le m} |\langle a_i, u\rangle| \cdot \sum_{i=1}^m |\ell'''(\langle a_i, x\rangle)|\langle a_i, h\rangle^2 \qquad (25.6) \\ &\le \|u\|_B \cdot \sum_{i=1}^m |\ell'''(\langle a_i, x\rangle)|\langle a_i, h\rangle^2, \end{aligned}$$

where we estimated the ℓ_∞-norm by the Euclidean one:

$$\max_{1\le i\le m} |\langle a_i, u\rangle| \le \sqrt{\sum_{i=1}^m \langle a_i, u\rangle^2} = \langle Bu, u\rangle^{1/2} =: \|u\|_B.$$

Thus, assuming that $\ell'''(t) \le L_2$, for all $t \in \mathbb{R}$, we get the following uniform bound on the third derivative:

$$|D^3 f(x)[h,h,u]| \le L_2\|u\|_B \cdot \sum_{i=1}^m \langle a_i, h\rangle^2 = L_2 \cdot \|u\|_B \cdot \|h\|_B^2.$$

The last inequality implies that the Hessian of f is Lipschitz continuous. Hence, we can apply the Cubic Newton method for this problem, that possesses the following global complexity, on convex functions (see exam questions):

$$O\Big(\Big[\tfrac{L_2 D^3}{\varepsilon}\Big]^{1/2}\Big), \qquad (25.7)$$

where $D \ge \|x_0 - x^\star\|_B$ is the size of the initial sublevel set measured in $\|\cdot\|_B$ norm:

$$D := \max\Big\{\|x - x^\star\|_B \;:\; f(x) \le f(x_0)\Big\}. \qquad (25.8)$$

We see that the complexity of the cubic Newton (25.7) is better than (25.5) of the gradient methods, in terms of the dependence on ε.

It is possible to accelerate the cubic Newton further, which we discuss in the next lecture.

Let us consider the following popular examples of the loss function.

Example 25.1.1 (Logistic Loss)**.**

$$\ell(t) = \ln(1 + e^t),$$

we have

$$L_1 = \tfrac{1}{4}, \qquad L_2 = \tfrac{1}{6\sqrt{3}}.$$

However, computing these constants, we may observe the following interesting relationship:

$$\ell'''(t) = \ell''(t) \cdot (1 - 2\ell'(t)) = \ell''(t) \cdot (1 - \tfrac{2}{1+e^{-t}}).$$

Hence, for logistic loss, it holds

$$|\ell'''(t)| \le \ell''(t), \qquad t \in \mathbb{R}. \qquad (25.9)$$

Example 25.1.2 (Exponential Loss)**.**

$$\ell(t) = e^t.$$

Note that $L_1 = L_2 = +\infty$ (globally), while (25.9) is satisfied as an exact equation.

Using (25.9) in our computations, we obtain:

$$|D^3 f(x)[h,h,u]| \overset{(25.6),(25.9)}{\leq} \|u\|_B \sum_{i=1}^m \ell''(\langle a_i, x\rangle)\langle a_i, h\rangle^2 \;=\; \|u\|_B \cdot \|h\|_x^2,$$

where $\|\cdot\|_x$ is the local norm induced by the Hessian (25.3). These observations motivate our next definition.

25.1.2 Definition of Quasi-Self-Concordant Functions

We consider a differentiable convex function $f : Q \to \mathbb{R}$, where $Q \subseteq \mathbb{R}^n$ is an open convex set. Without loss of generality, we can assume that $\nabla^2 f(x) \succ 0$ everywhere on Q. As usual, we denote by

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

the local norm at $x \in Q$ induced by the Hessian, and by $\|h\|$ we denote a fixed global norm. The main example for us is when the global norm is induced by a fixed positive definite operator $B = B^\top \succ 0$:

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

Definition 25.1.1. We say that a function $f : Q \to \mathbb{R}$ is *quasi-self-concordant* with constant $M \geq 0$, if

$$D^3 f(x)[h,h,u] \;\leq\; M\|h\|_x^2 \|u\|, \qquad \forall h, u \in \mathbb{R}^n, x \in Q. \tag{25.10}$$

The difference from classic self-concordant functions (see Lecture 18) is that we replace the local norm for u with the global fixed norm in the right-hand side of (25.10). Hence, the new problem class is no longer affine-invariant (the parameter M changes, if we change the coordinate system).

This definition can be seen as an intermediate problem class between self-concordant functions and the functions with Lipschitz continuous Hessians.

We see that logistic and exponential regression objectives satisfy assumption (25.10) with $\boxed{M = 1}$ by choosing the matrix $B = A^\top A$ as in the previous section, where A is the matrix representing input data, or $\boxed{M = \|A\|}$ when $B := I$. It is possible to show that they are not self-concordant in the classic sense:

Exercise 25.1.1. Show that both $\ell(t) = \ln(1 + e^t)$ and $\ell(t) = e^t$ are not self-concordant on $\mathbb{R}$, i.e., in each of these cases, there is no constant $M \geq 0$ such that

$$|\ell'''(t)| \;\leq\; M\big(\ell''(t)\big)^{3/2}, \qquad \forall t \in \mathbb{R}.$$

25.1.3 Main Properties

Let us take an arbitrary direction $h \in \mathbb{R}^n$ and consider how the local norm of u changes between two given points x and y. In particular, we look at the function

$$g(t) \;=\; \ln \|h\|^2_{x+t(y-x)} \;=\; \ln\langle \nabla^2 f(x + t(y-x))h, h\rangle, \qquad t \in [0,1].$$

Then,

$$\Big|g'(t)\Big| \;=\; \Big|\tfrac{D^3 f(x+t(y-x))[h]^2[y-x]}{\|h\|^2_{x+t(y-x)}}\Big| \overset{(25.10)}{\leq} M\|y-x\|.$$

Therefore,

$$\Big|\ln \tfrac{\|h\|_y^2}{\|h\|_x^2}\Big| \;=\; |g(1) - g(0)| \;=\; |\int_0^1 g'(t)dt| \;\leq\; M\|y-x\|.$$

Hence, taking the exponent:

$$\|h\|_x^2 e^{-M\|y-x\|} \;\leq\; \|h\|_y^2 \;\leq\; \|h\|_x^2 e^{M\|y-x\|}.$$

We have established the following main lemma, which is an analog of the Hessian stability for the quasi-self-concordant functions (compare with Lemma 19.1.1 from Lecture 19 on self-concordant functions, and with that one from Lecture 23 for the functions with Lipschitz Hessian):

Lemma 25.1.3. *For any $x, y \in \mathbb{R}^n$:*

$$\nabla^2 f(x) e^{-M\|y-x\|} \;\preceq\; \nabla^2 f(y) \;\preceq\; \nabla^2 f(x) e^{M\|y-x\|}. \tag{25.11}$$

Let us derive a consequence of our definition for the approximation of the gradient norm. For any direction $h \in \mathbb{R}^n$, s.t. $\|h\| \leq 1$, we have, using the Taylor theorem:

$$\begin{aligned}
\langle \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x), h\rangle &= \int_0^1 (1-t) D^3 f(x + t(y-x))[y-x]^2[h]dt \\
&\overset{(25.10)}{\leq} M\|h\| \cdot \int_0^1 (1-t)\|y-x\|^2_{x+t(y-x)} dt \\
&\overset{(25.11)}{\leq} M\|y-x\|_x^2 \cdot \int_0^1 (1-t) e^{tM\|y-x\|} d\tau \\
&\equiv M\|y-x\|_x^2 \cdot \varphi(M\|y-x\|).
\end{aligned}$$

By computing the integral, we obtain the following useful bound on the linear approximation of the gradient.

Lemma 25.1.4. *For any $x, y \in \mathbb{R}^n$:*

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\|_* \;\leq\; M\|y-x\|_x^2 \cdot \varphi(M\|y-x\|), \tag{25.12}$$

where $\varphi(t) := \frac{e^t - t - 1}{t^2} \geq 0$ is a monotone convex function (see Fig. 25.1, left).

Integrating (25.12) once more yields global upper and lower approximation bounds for an objective function (see Fig. 25.1, right).

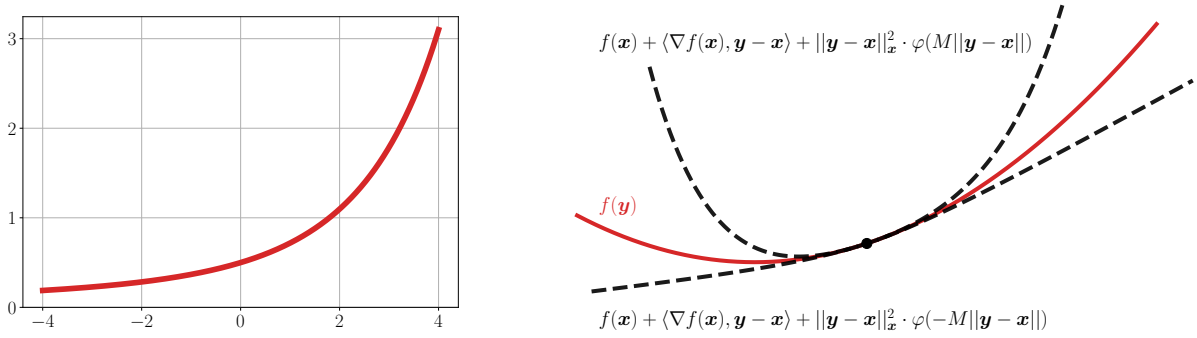

Figure 25.1: **Left:** The graph of $\varphi(t) = \frac{e^t - t - 1}{t^2}$ from the bound (25.12). **Right:** Global upper and lower models of a quasi-self-concordant function.

25.2 Gradient Regularization of Newton's Method

The cubic regularization of Newton's method is a very powerful approach as it works both for convex and non-convex functions, achieving superior rates to those of the first-order methods.

However, when solving convex problems, we can replace the nonlinear cubic subproblem by the quadratic regularization, which is easier to implement as each step requires solving only one linear system.

Recall that one step of the cubic Newton method with the regularization parameter $L \geq 0$ can be written in the following form (see Lecture 23):

$$x^+ \;=\; x - \left(\nabla^2 f(x) + \tfrac{Lr^\star}{2} B\right)^{-1} \nabla f(x), \tag{25.13}$$

where $r^\star$ is the solution of the nonlinear univariate equation, and we have $r^\star \approx \sqrt{\frac{1}{L}\|\nabla f(x^+)\|_*}$.

The idea of the *gradient regularization* is to replace the implicit regularization coefficient $r^\star$ with the current gradient norm $\|\nabla f(x)\|_*$, which can be easily computed at the current point x. Indeed, using the fact that $\nabla^2 f(x) \succeq 0$ for convex objectives, we can bound the length of the step from (25.13), as follows:

$$r^\star \;:=\; \|x^+ - x\| \;\overset{(25.13)}{=}\; \left\|\left(\nabla^2 f(x) + \tfrac{Lr^\star}{2} B\right)^{-1} \nabla f(x)\right\| \;\leq\; \tfrac{2}{Lr^\star}\|\nabla f(x)\|_*. \tag{25.14}$$

Hence, we obtain the upper bound:

$$r^\star \;\overset{(25.14)}{\leq}\; \sqrt{\tfrac{2}{L}\|\nabla f(x)\|_*}, \tag{25.15}$$

which we can use in (25.13) instead of $r^\star$. It appears that such an approximation preserves the fast global rates of the cubic Newton method [6, 4], for convex functions.

In general, we can consider iterations of the form:

$$x^+ \;=\; x - \left(\nabla^2 f(x) + H\|\nabla f(x)\|_*^\alpha B\right)^{-1} \nabla f(x), \tag{25.16}$$

where $0 \leq \alpha \leq 1$ is some fixed power, and $H \geq 0$ is a regularization parameter. Then, $\alpha = 0$ implies that we regularize the Hessian by a constant matrix, while substituting the upper bound (25.15) into (25.13) corresponds to $\alpha = 1/2$ and $H = \sqrt{2L}$.

Among the possible powers, the most appealing is $\boxed{\alpha = 1}$, as it preserves the *local quadratic convergence* of Newton's method, to be shown in the following exercise. This is the choice that we will analyze further for quasi-self-concordant functions.

Exercise 25.2.1. Consider step (25.16) for $\alpha = 1$ and some fixed $H \geq 0$. Assume that the function f is strongly convex, with a Lipschitz continuous Hessian (with corresponding parameters μ and L). Show that

$$\|\nabla f(x^+)\|_* \;\leq\; \left(\tfrac{L}{2\mu^2} + \tfrac{H}{\mu}\right)\|\nabla f(x)\|_*^2.$$

Therefore, the method possesses local quadratic convergence. What will be the local rate of the method with arbitrary $0 \leq \alpha \leq 1$?

25.2.1 Regularization by Gradient Power

Using $\alpha = 1$, iteration (25.16) can be rewritten as the solution to the linear system:

$$\nabla f(x) + \nabla^2 f(x)(x^+ - x) + H\|\nabla f(x)\|_* B(x^+ - x) \;=\; 0. \tag{25.17}$$

Taking the inner product with $x^+ - x$ and rearranging the terms, we get

$$\|x^+ - x\|_x^2 + H\|\nabla f(x)\|_*\|x^+ - x\|^2 \;=\; \langle \nabla f(x), x - x^+\rangle \;\le\; \|\nabla f(x)\|_*\|x^+ - x\|. \tag{25.18}$$

Dropping either the first or the second term, which are nonnegative, we obtain the following bounds:

Lemma 25.2.1. *It holds:*

$$\|x^+ - x\| \;\le\; \tfrac{1}{H}. \tag{25.19}$$

and

$$\|x^+ - x\|_x^2 \;\le\; \|\nabla f(x)\|_*\|x^+ - x\|. \tag{25.20}$$

Consequently, by performing gradient regularization, we automatically ensure that the iterates remain within the ball of radius $\frac{1}{H}$ centered at x. Furthermore, by (25.20), we can also control the radius of the ball in the local norm (the the radius of Dikin's ellipsoid).

25.2.2 Progress of One Step

Now, let us fix for simplicity $\boxed{H := M}$ — so we choose the regularization parameter to be exactly the constant of quasi-self-concordance. We combine the optimality condition (25.17) with our bound on the gradient approximation (25.12). Denote $g := \|\nabla f(x)\|_*$ and $r = \|x^+ - x\|$. First, note that

$$\varphi(M\|x^+ - x\|) \;\le\; \varphi(\tfrac{M}{H}) \;=\; \varphi(1) \;=\; \rho = e - 2 \;\approx\; 0.718281828.$$

$$\begin{aligned}\|\nabla f(x^+) + MgB(x^+ - x)\|_* &\le M\|x^+ - x\|_x^2 \cdot \varphi(M\|x^+ - x\|) \\ &\le \rho M g\|x^+ - x\|.\end{aligned}$$

Squaring both sides, we obtain:

$$g_+^2 + (Mgr)^2 + 2Mg\langle \nabla f(x^+), x^+ - x\rangle \;\le\; \rho^2 (Mgr)^2, \tag{25.21}$$

where $g_+ := \|\nabla f(x_+)\|_*$. Using the fact that $\rho < 1$, we obtain the following progress of one iteration.

Theorem 25.2.2. *For one Newton's step with gradient regularization, we have:*

$$f(x) - f(x^+) \;\ge\; \langle \nabla f(x^+), x - x^+\rangle \overset{(25.21)}{\ge} \tfrac{1}{2Mg} g_+^2 \;=\; \tfrac{1}{2M}\Big(\tfrac{\|\nabla f(x_+)\|_*}{\|\nabla f(x)\|_*}\Big)^2 \|\nabla f(x)\|_* \tag{25.22}$$

25.2.3 Global Linear Rate

Let us derive the rate of convergence from (25.22). We perform the following simple iterations:

$$x_{k+1} \;=\; x_k - \Bigl(\nabla^2 f(x_k) + M\|\nabla f(x_k)\|_* B\Bigr)^{-1} \nabla f(x_k), \qquad k \geq 0, \tag{25.23}$$

starting from an arbitrary initialization $x_0 \in \mathbb{R}^n$.

Denote the functional residual as $F_k := f(x_k) - f^\star$ and the gradient norm as $g_k := \|\nabla f(x_k)\|_*$. By convexity, we have

$$g_k \;\geq\; \tfrac{F_k}{D}, \tag{25.24}$$

where D is the diameter of the initial sublevel set as in (25.8). Substituting this into the progress of one step, we get:

$$F_k - F_{k+1} \;\overset{(25.22)}{\geq}\; \tfrac{1}{2M}\Bigl(\tfrac{g_{k+1}}{g_k}\Bigr)^2 g_k \;\overset{(25.24)}{\geq}\; \tfrac{1}{2MD}\Bigl(\tfrac{g_{k+1}}{g_k}\Bigr)^2 F_k. \tag{25.25}$$

It remains to derive the convergence rate from the recurrence (25.25). Rearranging the terms in (25.25), we see that

$$F_{k+1} \;\leq\; \Bigl[1 - \tfrac{1}{2MD}\Bigl(\tfrac{g_{k+1}}{g_k}\Bigr)^2\Bigr] \cdot F_k \;\approx\; \Bigl[1 - \tfrac{1}{2MD}\Bigr] \cdot F_k.$$

Thus, we can expect a linear rate of decrease for the sequence F_k, which suggests that the *appropriate quantity to telescope*[1] is $\ln(F_k)$.

We know that $\ln(a)$ is a concave function. Hence, for any $a, b > 0$:

$$\ln(a) \;\leq\; \ln(b) + \tfrac{1}{b}(a-b) \qquad \Leftrightarrow \qquad \ln(b) - \ln(a) \;\geq\; \tfrac{1}{b}(b-a), \tag{25.26}$$

Therefore,

$$\ln(F_k) - \ln(F_{k+1}) \;\overset{(25.26)}{\geq}\; \tfrac{F_k - F_{k+1}}{F_k} \;\overset{(25.25)}{\geq}\; \tfrac{1}{2MD}\Bigl(\tfrac{g_{k+1}}{g_k}\Bigr)^2. \tag{25.27}$$

Telescoping this bound, and using the inequality between arithmetic and geometric means (that is, Jensen's inequality for concavity of the logarithm), we get:

$$\begin{aligned}
\ln \tfrac{F_0}{F_k} \;&\overset{(25.27)}{\geq}\; \tfrac{k}{2MD} \cdot \tfrac{1}{k} \sum_{i=0}^{k-1} \Bigl[\tfrac{g_{i+1}}{g_i}\Bigr]^2 \;\geq\; \tfrac{k}{2MD} \cdot \Bigl[\prod_{i=0}^{k-1} \tfrac{g_{i+1}}{g_i}\Bigr]^{2/k} \\
&=\; \tfrac{k}{2MD} \cdot \Bigl[\tfrac{g_k}{g_0}\Bigr]^{2/k} \;=\; \tfrac{k}{2MD} \cdot \exp\Bigl(\tfrac{2}{k} \ln \tfrac{g_k}{g_0}\Bigr) \\
&\overset{(*)}{\geq}\; \tfrac{k}{2MD} \cdot \Bigl(1 + \tfrac{2}{k} \ln \tfrac{g_k}{g_0}\Bigr) \;\overset{(25.24)}{\geq}\; \tfrac{k}{2MD} \cdot \Bigl(1 + \tfrac{2}{k} \ln \tfrac{F_k}{g_0 D}\Bigr),
\end{aligned} \tag{25.28}$$

where in $(*)$ we used that $e^t \geq 1 + t$ for all $t \in \mathbb{R}$, which follows from convexity of e^t.

Consider two cases.

[1] Note that in the continuous-time case, the recurrence (25.25) takes the form $-\dot{F}_t \geq c \cdot F_t$ for a constant $c \geq 0$, which can be integrated to:

$$\ln \tfrac{F_0}{F_t} \;=\; \ln F_0 - \ln F_t \;=\; \int_0^t \tfrac{d}{dt}\bigl[-\ln F_t\bigr] \;=\; \int_0^t -\tfrac{\dot{F}_t}{F_t} dt \;\geq\; ct \qquad \Rightarrow \qquad F_t = O(F_0 e^{-ct}).$$

Integrating in continuous time corresponds to telescoping discrete sequences.

1. Either $\frac{2}{k} \ln \frac{F_k}{g_0 D} \leq -\frac{1}{2}$, which is equivalent to the very fast rate with constant factor:

$$F_k \;\leq\; \exp(-k/4) g_0 D.$$

2. Otherwise, $\frac{2}{k} \ln \frac{F_k}{g_0 D} \geq -\frac{1}{2}$. Substituting this bound into (25.28), gives

$$\ln \tfrac{F_0}{F_k} \;\geq\; \tfrac{k}{4MD} \qquad \Leftrightarrow \qquad F_k \;\leq\; \exp\Bigl(-\tfrac{k}{4MD}\Bigr) F_0.$$

Finally, we combine these two bounds together to obtain the following convergence rate.

Theorem 25.2.3. *For iterations of Newton's method with gradient regularization* (25.23)*, we have the global linear rate:*

$$f(x_k) - f^\star \;\leq\; \exp\Bigl(-\tfrac{k}{4MD}\Bigr)(f(x_0) - f^\star) + \exp\Bigl(-\tfrac{k}{4}\Bigr) g_0 D. \tag{25.29}$$

Therefore, in order to obtain $f(x_k) - f^\star \leq \varepsilon$ it is enough to perform the following number of iterations (second-order oracle calls):

$$k \;=\; O\Bigl(MD \ln \tfrac{F_0}{\varepsilon} + \ln \tfrac{g_0 D}{\varepsilon}\Bigr). \tag{25.30}$$

To establish (25.29), we did not use any additional assumptions, such as strong or uniform convexity, other than our main assumption of quasi-self-concordance (25.10).

The complexity bound (25.30) is superior to those of the gradient methods (25.5) and the cubically regularized Newton method (25.7) in terms of the final dependence on the target accuracy $\varepsilon > 0$. Note that in each of these situations, we are discussing not only different methods but, more importantly, *different problem classes*. For the basic (non-accelerated) methods, we have the following complexity picture:

- *Convex functions with Lipschitz gradient:* $O(1/\varepsilon)$
- *Convex functions with Lipschitz Hessian:* $O(1/\varepsilon^{1/2})$
- *Quasi-self-concordant functions:* $O(\ln \frac{1}{\varepsilon})$

At the same time, a single objective function can belong to multiple problem classes simultaneously. Therefore, for a given problem, we are primarily interested in the best possible convergence rate among the available options. For example, for training logistic regression, the quasi-self-concordant framework appears to provide the best global complexity (25.30) among those considered.

While we used a constant choice for the regularization parameter in (25.23), which requires knowing the constant of quasi-self-concordance, we can instead perform a simple adaptive search. This is analogous to the adaptive search used in gradient methods. Such an adaptive search will ensure sufficient progress (25.22) at each iteration. In fact, it was shown in [3] that employing the adaptive search allows the Newton method with gradient regularization to *automatically* achieve the best convergence rate among all the problem classes listed above, yielding *super-universal* guarantees.

Literature

Quasi-self-concordant functions were introduced in [1] and subsequently studied in [8, 5, 2]. The gradient regularization of Newton's method was considered in [7, 9, 6, 4, 3]. In these notes we followed the presentation from [2].

[1] Francis Bach. Self-concordant analysis for logistic regression. 2010.

[2] Nikita Doikov. Minimizing quasi-self-concordant functions by gradient regularization of Newton method. *Mathematical Programming*, pages 1–39, 2025.

[3] Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-universal regularized Newton method. *SIAM Journal on Optimization*, 34(1):27–56, 2024.

[4] Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with Bregman distances. *Mathematical programming*, 204(1):1–25, 2024.

[5] Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv preprint arXiv:1806.00413*, 2018.

[6] Konstantin Mishchenko. Regularized Newton method with global $\mathcal{O}(1/k^2)$ convergence. *SIAM Journal on Optimization*, 33(3):1440–1462, 2023.

[7] Roman A Polyak. Regularized Newton method for unconstrained convex optimization. *Mathematical programming*, 120:125–145, 2009.

[8] Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for Newton-type methods. *Mathematical Programming*, 178(1-2):145–213, 2019.

[9] Kenji Ueda and Nobuo Yamashita. A regularized Newton method without line search for unconstrained optimization. *Technical Report*, 2009.