Lecture 6

W. Hebisch

March 29, 2023

1 Unconstrained optimization

1.1 Unconstrained optimization, rate of convergence

Now we come to lower bound.

We will say that a method is first order method if

$$x_{i+1} \in x_0 + \operatorname{lin}\{\nabla f(x_0), \ldots, \nabla f(x_i)\}.$$

Intuitively, the condition above means that we use only information obtained from first derivatives to choose direction. Clearly gradient descent is a first order method, but we will see that most methods that we study are first order methods.

Lemma 1.1 *For any k, $1 \le k \le \frac{n-1}{2}$, and any $x_0 \in \mathbb{R}^n$ there exists a convex function $f : \mathbb{R}^n \to \mathbb{R}$ such that gradient of f is Lipschitz continuous with constant M and for any first-order method we have*

$$f(x_k) - f(x_\infty) \ge \frac{3M\|x_0 - x_\infty\|^2}{32(k+1)^2},$$

$$\|x_k - x_\infty\|^2 \ge \frac{1}{8}\|x_0 - x_\infty\|^2$$

where x_∞ is the minimum of f.

There is a method (Nesterow acceleration) for which we have corresponding upper bound.

Proof. To prove the lemma we need to construct appropriate couterexample. It is enough to do this for $x_0 = 0$.

For $s \in \mathbb{R}^n$ we define

$$f(s) = \frac{1}{2}\left(s_1^2 + s_n^2 + \sum_{i=1}^{n-1}(s_{i+1} - s_i)^2\right) - s_1.$$

Clearly $\nabla^2 f$ is strictly positive definite. One can check that $\nabla f(s) = 0$ when

$$s_i = 1 - \frac{i}{n+1}$$

so optimal value is $\frac{1}{2}(-1+\frac{1}{n+1})$.

Let V_i be subspace of $\mathbb{R}^n$ consisting from vectors such that only first i coordinates are nonzero ($V_0 = \{0\}$). One can check that when $x_i \in V_i$, then $\nabla f(x_i) \in V_{i+1}$ so for any first order method starting at $x_0 = 0$ we have $x_i \in V_i$.

At best first order metod will give optimal value on V_i. However, f restricted to V_i looks exactly like f with $n = i$, so optimal value is $\frac{1}{2}(-1+\frac{1}{i+1})$ and therefore error

$$E(x_i) = \frac{1}{2}(-1+\frac{1}{i+1}) - \frac{1}{2}(-1+\frac{1}{n+1})$$

$$= \frac{1}{2}(\frac{1}{i+1} - \frac{1}{n+1})$$

and in particular when $2(i+1) \leq n+1$ then

$$E(x_i) \geq \frac{1}{4}\frac{1}{i+1}.$$

Also, note that

$$\|x_i\|^2 = \sum_{j=1}^{i}(1-\frac{j}{i+1})^2 = \sum_{j=1}^{i}\frac{j^2}{(i+1)^2}$$

$$\leq \sum_{j=1}^{i}\frac{(j+1)^3-j^3}{3(i+1)^2} = \frac{(i+1)^3}{3(i+1)^2} - \frac{1}{3(i+1)^2} \leq \frac{i+1}{3}$$

so

$$E(x_i) \geq O(\frac{1}{(i+1)^2})\|x_0 - x_\infty\|^2$$

Moreover,

$$\|x_i - x_\infty\|^2 \geq c\|x_0 - x_\infty\|^2$$

□

Previous lemma means that even for quite requlars convex functions we can not expect very fast convergence. In fact, there is already problem with convex quadratic functions.

Above we have difficulty because f is convex, but second derviative may be very small in some directions (and relatively large in other directions). So we need lower bound on second derivative. To use similar notation as in following lectures we introduce here Hessian:

$$\langle \nabla^2 f(x)h_1, h_2\rangle = f''(x)(h_1, h_2)$$

that is for fixed x right hand side is a (symmetric) quadratic form in h, while on left hand side $\nabla^2 f(x)$ is a linear operator uniqely defined by the equality above.

Technically, instead of lower bound on $\nabla^2 f$ it is more elegant to use following condition.

We say that f is *strongly convex with constant* m when $f - \frac{1}{2}m\|x\|^2$ is convex.

This condition holds when f is twice differentiable and all eigenvalues of $\nabla^2 f$ are bounded from below by m. However, as written above condition does not require existence of second deriviative.

Lemma 1.2 *When f is strongly convex with constant m, that is $f - \frac{1}{2}m\|x\|^2$ is convex, gradient of f is Lipschitz continuous with constant M, then for gradient descent with constant step size $\alpha = \frac{2}{M+m}$ we have*

$$\|x_i - x_\infty\| \leq C^i \|x_0 - x_\infty\|$$

where

$$C = \frac{M-m}{M+m}$$

and x_∞ is a minimal point.

Proof: We will represent ∇f using $\nabla^2 f$ (which is possible under our assumptions, in particular $\nabla^2 f$ exists almost everywhere):

$$\nabla f(x_i) = \nabla f(x_i) - \nabla f(x_\infty) = \int_0^1 \nabla^2 f(x + th_i) h_i$$

where $h_i = x_i - x_\infty$. So

$$h_{i+1} = x_{i+1} - x_\infty = x_i - x_\infty - \alpha \nabla f(x_i)$$

$$= (I - \alpha \int_0^1 \nabla^2 f(x + th_i)) h_i = A h_i$$

By assumption $mI \leq \nabla^2 f(x + th_i) \leq MI$, so

$$\frac{2m}{M+m} I \leq \alpha \nabla^2 f(x + th_i) \leq \frac{2M}{M+m} I.$$

So

$$-\frac{M-m}{M+m} I \leq A \leq \frac{M-m}{M+m} I$$

and since A is symmetric

$$\|A\| \leq \frac{M-m}{M+m}.$$

Consequently

$$\|x_{i+1} - x_\infty\| \leq \|A\| \|x_i - x_\infty\| \leq C \|x_i - x_\infty\|$$

and claim follows by induction. □

Remark: Again, slightly enlarging C we get result for steps smaller than $\frac{2}{M+m}$ and for slightly larger steps. But the proof breaks down when steps are much larger.

We also get result for improvement of goal function:

Lemma 1.3 *With assumptions as above and step* $\alpha = \frac{1}{M}$

$$f(x_{i+1}) - f(x_\infty) \leq C(f(x_i) - f(x_\infty))$$

where $C = 1 - \frac{m}{M}$.

Proof: In previous lecture we obtained decay estimate

$$f(x_{i+1}) \leq f(x_i) - \frac{1}{2M}\|\nabla f(x_i)\|^2$$

so

$$f(x_{i+1}) - f(x_\infty) \leq f(x_i) - f(x_\infty) - \frac{1}{2M}\|\nabla f(x_i)\|^2$$

By strong convexity

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|y - x\|^2.$$

Computing derivative we see that $z = x - \frac{1}{m}\nabla f(x)$ minimizes right hand side (with fixed x and variable y).

So

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|y - x\|^2 \\ &\geq f(x) + \langle \nabla f(x), z - x\rangle + \frac{m}{2}\|z - x\|^2 \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \end{aligned}$$

Using $y = x_\infty$ and $x = x_i$ we get

$$\frac{1}{2m}\|\nabla f(x_i)\|^2 \geq f(x_i) - f(x_\infty)$$

Plugging the above into decay estimate we get

$$\begin{aligned} f(x_{i+1}) - f(x_\infty) &\leq f(x_i) - f(x_\infty) - \frac{1}{2M}2m(f(x_i) - f(x_\infty)) \\ &= C(f(x_i) - f(x_\infty)) \end{aligned}$$

with $C = 1 - \frac{m}{M}$ which gives the claim. □

Remark: clearly exact line search is as good or better than this. Armijo's rule gives at least fraction of decay of fixed step, so we get similar result with slightly larger C.

Remark: When $\frac{m}{M}$ is reasonably big this is much better than result without strong convexity: we need number of steps that grows linearly with accuracy while $\frac{1}{\epsilon}$ is exponential in number of bits.

What to do when $\frac{m}{M}$ is very small? It may happen that $\frac{m}{M}$ is small due to bad scaling, this happened in quadratic example from previous lecture. Simple

diagonal scaling sometimes helps, but rotating bad example we see that in general we may need to rescale by arbitrary linear mapping A. In a sense optimal rescaling would give

$$\|Ah\|^2 = \langle \nabla^2 f(x)h, h\rangle$$

We can not get this for all x simultaneously, but can do for single point x_i. Namely, we define new scalar product as $\langle h, h\rangle_x = \langle \nabla^2 f(x)h, h\rangle$. Then steepest descent direction is given by Newton formula

$$(\nabla^2 f(x_i))^{-1}\nabla f(x_i).$$

1.2 Newton method

When $x_{i+1} = x_i - (\nabla^2 f(x_i))^{-1}\nabla f(x_i)$ we say about pure Newton method.

To derive Newton formula we find steepest descent direction at x minimizing

$$\frac{\langle \nabla f(x), h\rangle}{\langle h, h\rangle_x^{1/2}} = \frac{\langle \nabla f(x), h\rangle}{\langle \nabla f^2(x)h, h\rangle^{1/2}}.$$

Computing derivative with respect to h we get

$$\nabla f(x)\langle \nabla^2 f(x)h, h\rangle^{1/2} - \langle \nabla f(x), h\rangle \nabla^2 f(x)h$$

as numerator. Comparing this to 0 gives equation

$$\langle \nabla^2 f(x)h, h\rangle^{1/2}\nabla f(x) = \langle \nabla f(x), h\rangle \nabla^2 f(x)h$$

that is

$$h = \frac{\langle \nabla^2 f(x)h, h\rangle^{1/2}}{\langle \nabla f(x), h\rangle}(\nabla^2 f(x))^{-1}\nabla f(x).$$

Note that

$$\frac{\langle \nabla^2 f(x)h, h\rangle^{1/2}}{\langle \nabla f(x), h\rangle}$$

is just a negative constant, so in fact we get $(\nabla^2 f(x))^{-1}\nabla f(x)$ as the steepest descent direction.

Classically, Newton method was obtained looking at quadratic approximation to $f(x+h)$:

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x)h, h\rangle + o(\|h\|^2).$$

Minimizing quadratic function of h on the right hand side gives

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

So we get iterative algorithm

$$x_{i+1} = x_i - \alpha(\nabla^2 f(x))^{-1}\nabla f(x).$$

$\alpha = 1$ is pure Newton method, when α may be smaller than 1 we have damped Newton method.

Newton method for optimization is closely related to Newton method for equation solving, namely Newton method for optimization of f produces the same approximations as Newton method for solving equation $\nabla f(x) = 0$. When solving equations it is easy to see that pure Newton method may diverge, that is approximations converge to a cycle or are chaotic. The same may happen when using pure Newton method for optimization.

Consider now $g(y) = f(Ay + b)$ where A is an invertible matrix. We have

$$\langle \nabla g(y), h\rangle = \langle \nabla f(Ay + b), Ah\rangle,$$
$$g''(y)(h, h) = f''(Ay + b)(Ah, Ah)$$

so

$$\nabla g(y) = A^T \nabla f(Ay + b),$$
$$\nabla^2 g(y)h = A^T(\nabla^2 f)(Ay + b)Ah$$

and

$$(\nabla^2 g(y))^{-1}\nabla g(y) = A^{-1}(\nabla^2 f(Ay + b))^{-1}(A^T)^{-1}A^T\nabla f(Ay + b)$$
$$= A^{-1}\nabla^2 f(Ay + b))^{-1}\nabla f(Ay + b)$$

Rewriting, we get

$$A(\nabla^2 g(y))^{-1}\nabla g(y) = (\nabla^2 f(Ay + b))^{-1}\nabla f(Ay + b).$$

Let y_i be approximations produced by Newton method applied to g and let $w_i = (\nabla^2 g(y_i))^{-1}\nabla g(y_i)$ be corresponding search directions. Put $x_i = Ay_i + b$. Our result shows that starting Newton method for f at x_i we get $h_i = Aw_i$ as as search direction. Clearly

$$g(y_i + \alpha w_i) = f(A(y_i + \alpha w_i) + b) = f(Ay_i + b + \alpha Aw_i) = f(x_i + \alpha h_i)$$

so we use the same function of α during line search, so we will get the same α_i. Consequently, starting Newton method for f at x_0 we will get x_i at step i.

In other words, Newton method is invariant under affine transformations: changing variables in affine way changes approximations produced by Newton method in the same way. This is quite different than gradient descent, where change of variables plays much more role.

Note: to have true invariance we should have invariant stopping criterion. Good criterion is given by smallness of step h_i using our scalar product

$$\langle h_i, h_i\rangle_{x_i} = \langle \nabla^2 f(x_i)h_i, h_i\rangle = -\langle \nabla^2 f(x_i)(\nabla^2 f(x_i))^{-1}\nabla f(x_i), h_i\rangle$$
$$= -\langle \nabla f(x_i), h_i\rangle.$$

Using our formulas we have

$$-\langle \nabla g(y_i), w_i\rangle = -\langle A^T\nabla f(Ay_i + b), A^{-1}h_i\rangle = -\langle \nabla f(x_i), h_i\rangle$$

so requirement $-\langle \nabla f(x_i), h_i\rangle < \varepsilon$ gives invariant stopping criterion.

1.3 Local convergence

To analyze local convergence, we start with Newton method for equation solving. Let $g : \mathbb{R}^n \to \mathbb{R}^n$. Given x_i we define

$$x_{i+1} = x_i - g'(x_i)^{-1} g(x_i)$$

Assume that x_∞ is a solution, that is $g(x_\infty) = 0$.

Let $B_r = \{x : \|x - x_\infty\| < r\}$ be ball of radius r around x_∞.

Lemma 1.4 *Assume that g' is Lipschitz continuous with constant L on B_r and $\|g'^{-1}(x)\| \leq m$. If $x_i \in B_r$, then*

$$\|x_{i+1} - x_\infty\| \leq \frac{Lm}{2} \|x_i - x_\infty\|^2$$

If additionally $rLm \leq 2$, then $x_{i+1} \in B_r$. Also, when $rLm \leq 2$, it is enough to assume that $\|g'^{-1}(x)\| \leq m$ for $x \in B_r$.

Proof: Put $\delta_i = x_i - x_\infty$. We compute

$$0 = g(x_\infty) = g(x_i) - \int_0^1 g'(x_i - t\delta_i)\delta_i dt$$

$$= g(x_i) - g'(x_i)\delta_i - \int_0^1 (g'(x_i - t\delta_i) - g'(x_i))\delta_i dt$$

so

$$\|g(x_i) - g'(x_i)\delta_i\| \leq \|\delta_i\| \int_0^1 \|g'(x_i - t\delta_i) - g'(x_i)\| dt$$

$$\leq \|\delta_i\| \int_0^1 L\|\delta_i\| t dt = \frac{L}{2}\|\delta_i\|^2$$

Now

$$x_{i+1} - x_\infty = x_i - g'(x_i)^{-1} g(x_i) - x_\infty = g'(x_i)^{-1}(g'(x_i)(x_i - x_\infty) - g(x_i))$$

$$= g'(x_i)^{-1}(g'(x_i)\delta_i - g(x_i))$$

so

$$\|x_{i+1} - x_\infty\| \leq \|g'(x_i)^{-1}\| \|g'(x_i)\delta_i - g(x_i)\|$$

$$\leq m\frac{L}{2}\|\delta_i\|^2 = \frac{mL}{2}\|x_i - x_\infty\|^2$$

which is the required estimate.

Since $x_i \in B_r$ we have $\|x_i - x_\infty\| < r$ and if additional assumption is satisfied we have

$$1 \geq \frac{rmL}{2} > \frac{mL\|x_i - x_\infty\|}{2}$$

so

$$\|x_{i+1} - x_\infty\| \leq \frac{mL\|x_i - x_\infty\|}{2} \|x_i - x_\infty\| < \|x_i - x_\infty\|$$

and $\|x_{i+1} - x_\infty\| < \|x_i - x_\infty\| < r$, hence $x_{i+1} \in B_r$.

Finally, we used estimate on $\|g'^{-1}\|$ only for x in line segment joining x_i and x_∞. When $rmL < 2$ this line segment is inside B_r. □

Assuming that g is regular enough and $g'(x_\infty)$ is invertible from the lemma we see that once Newton method is sufficiently close to solution it will converge, moreover, each step approximately doubles accuracy (number of significant bits). So, once we are close to solution 5 or 6 steps usually is enough to get full machine accuracy.

Under reasonable global assumptions it is possible to find out if we are close enough to solution:

Lemma 1.5 *Assume that g' is Lipschitz continuous with constant L and put $m = \|g'(x_0)^{-1}\|$. If*

$$\|g(x_0)\| \leq \frac{3}{16Lm^2},$$

then Newton method starting at x_0 converges to x_∞ and

$$\|x_0 - x_\infty\| \leq \frac{1}{4Lm}.$$

The result remains valid if we only assume that g' is Lipschitz continuous with constant L on ball centered at x_0 and radius $\frac{1}{2Lm}$.

Idea of the proof: let $A = g'(x_0)^{-1}$ and $\phi(x) = Ag(x)$. $\phi'(x) = Ag'(x)$, so $\phi'(x_0) = I$ (identity matrix) and ϕ' is Lipschitz continuous with constant Lm. As long as $\|x - x_0\| \leq \frac{1}{2Lm}$ we have $\|\phi'(x) - I\| \leq \frac{1}{2}$, so $\phi'(x)$ is invertible and $\|\phi'(x)\| \leq 2$. If $\|x_0 - x_\infty\| \leq \frac{1}{4Lm}$, then by previous lemma Newton method applied to ϕ is convergent and

$$\|x_0 - x_\infty\| \leq \|x_0 - x_1\| + \|x_1 - x_\infty\| \leq \|\phi(x_0)\| + Lm\|x_0 - x_\infty\|^2$$

$$\leq m\|g(x_0)\| + \frac{1}{4}\|x_0 - x_\infty\|$$

so

$$\frac{3}{4}\|x_0 - x_\infty\| \leq m\|g(x_0)\|$$

and

$$\|x_0 - x_\infty\| \leq \frac{4}{3}m\|g(x_0)\| \leq \frac{1}{4Lm}$$

To drop assumption $\|x_0 - x_\infty\| \leq \frac{1}{4Lm}$ consider equation $\phi(x) - t\phi(x_0) = 0$. For $t = 1$ this has solution in $K = \{x : \|x - x_0\| < \frac{1}{4Lm}\}$, namely x_0. By inverse function theorem set of t such that we have solution in K is open. By compactness set of t such that we have solution in $\bar{K}$ (closure of K) is closed. By previous estimate, for $t > 0$ solution must be in K, so set of t such that we have solution is K is nonempty open and closed subset of $(0, 1]$, so since interval is connected it is whole $(0, 1]$. By compactness for $t = 0$ we get solution such

that $\|x - x_0\| \leq \frac{1}{4Lm}$. □

We could restate the results for optimization, we will formulate appropriate lemma later.

1.4 Global convergence

Local convergence means that close to solution $\alpha_i = 1$ is good choice of step size. In fact, assuming regularity there is little motivation to use steps bigger than 1, but to get global convergence we sometimes need step smaller than 1. More precisely, when sufficient decay condition is violated we decrease step size.

In fact, local convergence implies that as long as we get decay we also get global convergence of gradient to zero.

However, there are two difficulties. First, in general (when $\nabla^2 f$ is not positive definite) Newton direction may fail to be decay direction. Second, local convergence assumes that $\nabla^2 f$ is invertible at stationary points. We can avoid both difficulties adding to $\nabla^2 f$ multiple of identity to make it positive definite. Unfortunately, it is hard to give some warranty for such method. Instead, we will assume strong convexity.

Assume that $mI \leq \nabla^2 f(x) \leq MI$ and $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L$. By strong convexity there is optimal point. Since $\nabla^2 f(x)$ is positive definite Newton direction is descent direction and we have global convergence.

For local convergence we have the following results:

Lemma 1.6 *If $\|\nabla f(x_i)\| < \frac{2m^2}{L}$, then pure Newton method starting at x_i is convergent and*

$$\frac{L}{2m^2}\|\nabla f(x_{i+1})\| \leq \left(\frac{L}{2m^2}\|\nabla f(x_i)\|\right)^2$$

Proof: We write $h_i = -(\nabla^2 f(x_i))^{-1}\nabla f(x_i)$ so $\nabla^2 f(x_i)h_i + \nabla f(x_i) = 0$ and

$$\nabla f(x_{i+1}) = \nabla f(x_i + h_i) - \nabla f(x_i) - \nabla^2 f(x_i)h_i$$

$$= \int_0^1 (\nabla^2 f(x_i + th_i) - \nabla^2 f(x_i))h_i dt$$

so

$$\|\nabla f(x_{i+1})\| \leq \int_0^1 L\|th_i\|\|h_i\|dt = \frac{L\|h_i\|^2}{2}$$

Since $\|\nabla^2 f(x_i)^{-1}\| \leq \frac{1}{m}$ we have $\|h_i\| \leq \frac{1}{m}\|\nabla f(x_i)\|$ so

$$\|\nabla f(x_{i+1})\| \leq \frac{L\|\nabla f(x_i)\|^2}{2m^2}$$

which gives bound on $\|\nabla f(x_{i+1})\|$. Under assumption this decreases which proves convergence. □

Recall Armijo's rule:

$$f(x_i + \alpha h_i) - f(x_i) \leq \rho\alpha\langle\nabla f(x_i), h_i\rangle.$$

Lemma 1.7 *If* $\|\nabla f(x_i)\| \leq \frac{3(1-2\rho)m^2}{L}$, *then* $\alpha = 1$ *is acceptable by Armijo's rule.*

Let $\gamma = \frac{\rho m}{\eta M^2}$

Lemma 1.8 *Assume* $\rho \leq \frac{1}{2}$. *If step size* α *in Newton method is selected starting from* $\alpha = 1$ *and dividing* α *by* η *as long as Armijo's condition is violated, then*

$$f(x_{i+1}) - f(x_i) \leq -\gamma\|\nabla f(x_i)\|^2$$

Proof: Put $\lambda = -\langle\nabla f(x_i), h_i\rangle = \langle\nabla^2 f(x)h_i, h_i\rangle$. By strong convexity $\langle h_i, h_i\rangle \leq \frac{1}{m}\lambda$. Using $\nabla^2 f(x) \leq MI$ we have

$$f(x_i + \alpha h_i) \leq f(x_i) + \alpha\langle\nabla f(x_i), h_i\rangle + \frac{M}{2}\alpha^2\|h_i\|^2$$

$$\leq f(x_i) - \alpha\lambda + \frac{M}{2m}\alpha^2\lambda.$$

Now we see that $\alpha = \frac{m}{M}$ satisfies Armijo's rule

$$f(x_i + \alpha h_i) \leq f(x_i) - \alpha\lambda + \frac{1}{2}\alpha\lambda$$

$$= f(x_i) - \frac{1}{2}\alpha\lambda$$

since $\rho \leq \frac{1}{2}$. Therefore we choose at least $\alpha = \frac{m}{\eta M}$ leading to decay

$$f(x_{i+1}) - f(x_i) \leq -\rho\alpha\lambda$$

$$\leq -\frac{\rho m}{\eta M}\lambda$$

Since $\lambda \leq \frac{1}{M}\|\nabla f(x_i)\|^2$ this gives

$$f(x_{i+1}) - f(x_i) \leq -\frac{\rho m}{\eta M^2}\|\nabla f(x_i)\|^2 = -\gamma\|\nabla f(x_i)\|^2.$$

□

The lemmas together imply global convergence of Newton method with $\rho < 1/2$: as long as $\|\nabla f(x_i)\|$ is big (so that local convergence does not apply) we get steady decay of value of f, in fact, putting $t = \min(3(1-2\rho), 1)\frac{m^2}{L}$ in at most

$$\frac{f(x_0) - f(x_\infty)}{\gamma t^2}$$

steps $\|\nabla f(x_i)\| \geq t$. But once $\|\nabla f(x_i)\| < t$ local convergence holds and we get any fixed accuracy in a fixed number of steps.

1.5 Further reading

Stephen Boyd, Lieven Vandenberghe, Convex Optimization, chapter 9.

David G. Luenberger, Yinyu Ye, Linear and Nonlinear Programming, chapters 7 and 8.

Yurii Nesterov, Introductory lectures on convex optimization, Springer 2004, chapter 1 (despite title more advanced than other texts).

Jorge Nocedal, Stephen J. Wright, Numerical Optimization, chapter 3.