16 Newton's method

Tuesday, October 27, 2020 2:23 AM

Finishing up convexity from last time

Def. 4.9/4.6 f has a minimum at $u \in E$ if $f(u) \le f(v) \quad \forall v \in E$.

strict minimum at $u \in E$ if $f(u) < f(v) \quad \forall v \in E - \{u\}$.

f has a minimum at $u \in U \subseteq E$ w.r.t. U if $f(u) \le f(v) \quad \forall v \in U$.

strict minimum at $u \in U \subseteq E$ w.r.t. U if $f(u) < f(v) \quad \forall v \in U - \{u\}$.

We sometimes stress that these are global minimums (vs. local)

Thm 4.5/4.11 Given any normed vector space E, let $U \subseteq E$ be nonempty and convex.

(1) For any convex function $J: U \to \mathbb{R}$, $\forall u \in U$, if J has a local minimum at u in U, then J has a (global) minimum at u in U.

(2) Any strict convex function $J: U \to \mathbb{R}$ has at most one minimum (in U). If it does have a minimum, then it is a strict minimum (in U).

(3) Let $J: \Omega \to \mathbb{R}$ be any function defined on an open subset $\Omega \subseteq E$ with $U \subseteq \Omega$ convex and assume J is convex on U. $\forall u \in U$, if $dJ(u)$ exists, then J has a minimum in u w.r.t. U iff

$$dJ(u)(v-u) \ge 0 \quad \forall v \in U.$$

(4) If the convex subset U in (3) is open, then the above condition is equivalent to $dJ(u) = 0$.

proof sketch: Repeated application of definitions and convexity.

Punchline is that for convex functions, local and global minima correspond.

Ex. 4.7/— Consider least squares solution to $Ax = b$.

i.e. Find $\arg\min_v \|Av - b\|_2^2$.

Consider $J(v) = \frac{1}{2}\|Av-b\|_2^2 - \frac{1}{2}\|b\|_2^2$, a quadratic function.

$$= \frac{1}{2}(Av-b)^T(Av-b) - \frac{1}{2}b^Tb$$

$$= \frac{1}{2}(v^TA^T - b^T)(Av-b) - \frac{1}{2}b^Tb$$

$$= \frac{1}{2}v^TA^TAv - v^TA^Tb$$

$\Rightarrow \nabla J(\;) = A^TA \;\; - A^Tb$

$\Rightarrow \quad dJ(u) = A^TAu - A^Tb.$

Since A^TA is pos. semidefinite, J is convex, then the last theorem implies that the (global) minima of J are solutions to $A^TAu - A^Tb = 0$

Recall SVD $A = VDU^T$ and $A^+ = UD^+V^T$. Let $u = A^+b$.

Then $UDV^TVDU^TUD^+V^Tb = UDV^Tb = A^Tb.$

Conclusion: We want to find zeros of $dJ: \Omega \to E'$, as that tells us a lot about minima. Next time, generalizations of Newton's method.

Newton's method

Given $f: \Omega \to Y$, where Ω is an open subset of a normed vector space X, and Y is a normed vector space, find

(1) Sufficient conditions to guarantee the existence of $a \in \Omega$ s.t. $f(a) = 0$.

(2) An algorithm for approximating such $a \in \Omega$. i.e. a sequence (x_k), $x_i \in \Omega$ s.t. $\lim_{i \to \infty} x_i = a$.

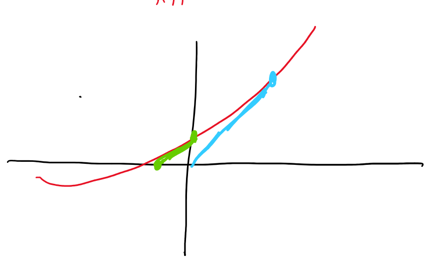
Recall: Newton's method.

Let $X = Y = \mathbb{R}$, $\Omega \subseteq \mathbb{R}$ open, $f: \Omega \to \mathbb{R}$.

Pick $x_0 \in \mathbb{R}$ "close enough" to a zero a s.t. $f(a) = 0$.

Define $x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)}$, for $k \geq 0$ provided $f'(x_k) \neq 0$.

Note that x_{k+1} is the 0 of the tangent line to $f(x)$ at x_k.



For arbitrary spaces: $f: \Omega \to Y$, $\Omega \subseteq X$, starting pt x_0.

Let $x_{k+1} = x_k - \left(f'(x_k)\right)^{-1}\left(f(x_k)\right)$

Need: Ω — existence of derivative — derivative to be invertible.

Ex. $f: \Omega \to \mathbb{R}^n$, $\Omega \subseteq \mathbb{R}^n$.

Need $f(a) = 0$, $a \in \mathbb{R}^n$.

Standard Newton's: $x_{k+1} = x_k - (f'(x_k))^{-1} f(x_k)$

Now let $\Delta x_k = x_{k+1} - x_k$, so $x_{k+1} = x_k + \Delta x_k$.

$$\Rightarrow \Delta x_k = -(f'(x_k)^{-1}) f(x_k)$$

$$\Rightarrow f'(x_k)\, \Delta x_k = -f(x_k)$$ $\Leftarrow$ $Ax = b$, solving linear system of equations

Variant 1: $(J(f)(x_k))\Delta x_k = -f(x_k)$ $\qquad J(f)(x_k) = \left(\frac{\partial f_i}{\partial x_j}(x_k)\right)_{i,j}$ Jacobian

Plain Newton's is expensive since we are computing a Jacobian and solving $Ax=b$ at every iteration

Variant 2: Use the same Jacobian for p consecutive steps

$$x_{k+1} = x_k - \left(f'(x_{\lfloor k/p \rfloor \cdot p})\right)^{-1}(f(x_k)).$$

Variant 3: Set $p = \infty$

$$x_{k+1} = x_k - (f'(x_0))^{-1}(f(x_k))$$

Variant 4: Replace $f'(x_0)$ entirely with an easy-to-invert A_0.

$$x_{k+1} = x_k - A_0^{-1} f(x_k)$$

Can further speed things up using LU factorization, or sometimes even $A_0 = I$.

Def. 5.1 If X and Y are normed vector spaces, and $f: \Omega \to Y$, $\Omega \subseteq X$ open, a *generalized Newton method* for finding 0's of f consists of:

(1) A sequence of families $(A_k(x))$ of linear isomorphisms from X to Y, for all $x \in \Omega$ and all integers $k \geq 0$.

(2) Some starting point $x_0 \in \Omega$.

(3) A sequence (x_k) of points of Ω defined by

$$x_{k+1} = x_k - (A_k(x_k))^{-1}(f(x_k)) \qquad k \geq 0.$$

(3) A sequence (x_k) of points [illegible]

$$x_{k+1} = x_k - \left(A_k(x_\ell)\right)^{-1} f(x_k), \quad k \geq 0,$$

where $\forall\, k \geq 0$, $\quad 0 \leq \ell \leq k$.

Or equivalently, $\quad A_k(x_\ell)(\Delta x_k) = -f(x_k)$

and $\quad x_{k+1} = x_k + \Delta x_k$.

Usually, $A_k(x)$ depends on f'.

Ex. 5.2 Let $f(X) = A - X^{-1}$.

Then $f'_X(Y) = X^{-1} Y X^{-1} \Rightarrow (f'_X)^{-1}(Y) = XYX$

Newton's method: $\quad X_{k+1} = X_k - (f'_{X_k})^{-1}(f(X_k)) = X_k - X_k(A - X_k^{-1})X_k$

$$= X_k(2I - AX_k).$$

Of course, Newton's method doesn't always converge. Can we give sufficient conditions?

(inspired by Newton - Kantorovich)

Thm 5.1 / 5.1 Let X be a Banach space, $f: \Omega \to Y$ differentiable on the open subset $\Omega \subseteq X$, and **assume** $\exists$ constants $r, M, \beta > 0$ s.t. if

$$B = \{x \in X \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then (1) $\displaystyle \sup_{k \geq 0} \sup_{x \in B} \|A_k^{-1}(x)\|_{\mathcal{L}(Y;X)} \leq M,$ (The inverse of the continuous linear map is bounded in norm)

(2) $\displaystyle \beta < 1$ and $\displaystyle \sup_{k \geq 0} \sup_{x' \in B} \|f'(x) - A_k(x')\|_{\mathcal{L}(Y;X)} \leq \frac{\beta}{M}$ (The difference of the map to the derivative is bounded in norm)

(3) $\displaystyle \|f(x_0)\| \leq \frac{r}{M}(1 - \beta).$ (A condition on the niceness of the initial guess)

Then the sequence (x_k) defined by

$$x_{k+1} = x_k - A_k^{-1}(x_\ell)(f(x_k)), \quad 0 \leq \ell \leq k$$

is entirely contained within B and converges to a, s.t. $f(a) = 0$, and a is the only zero of f in B. Furthermore, the convergence is

is entirely contained within ... ,
is the only zero of f in B. Furthermore, the convergence is geometric, so
$$\|x_k - a\| \le \frac{\|x_1 - x_0\|}{1-\beta} \cdot \beta^k .$$

proof. Lemma 1: For all $k \ge 1$,
$$\|x_k - x_{k-1}\| \le M \|f(x_{k-1})\|$$
$$\|x_k - x_0\| \le r \qquad (x_k \in B)$$
$$\|f(x_k)\| \le \frac{\beta}{M} \|x_k - x_{k-1}\|$$

proof. Induction on k.

- Base case: $k=1$.
$$x_1 = x_0 - A_0^{-1}(x_0)(f(x_0)) \qquad \text{by definition}$$

$$\Rightarrow \quad x_1 - x_0 = -A_0^{-1}(x_0)(f(x_0))$$

$$\Rightarrow \quad \|x_1 - x_0\| \le M \|f(x_0)\| \le r(1-\beta) \le r.$$

Also, $f(x_0) = -A_0(x_0)(x_1 - x_0)$ by rearranging the def. of x_1.
$$\Rightarrow \quad 0 = -f(x_0) - A_0(x_0)(x_1 - x_0)$$
$$\Rightarrow \quad f(x_1) = f(x_1) - f(x_0) - A_0(x_0)(x_1 - x_0).$$

By the MVT applied to $x \mapsto f(x) - A_0(x_0)(x)$, (Prop. 3.10)
$$\|f(x_1)\| \le \sup_{x \in B} \|f'(x) - A_0(x_0)\| \, \|x_1 - x_0\| \le \frac{\beta}{M} \|x_1 - x_0\|.$$

- Induction step: $k \ge 2$.
$$x_k - x_{k-1} = -A_{k-1}^{-1}(x_\ell)(f(x_{k-1})), \quad 0 \le \ell \le k-1 \qquad \text{by def.}$$
$$\Rightarrow \quad \|x_k - x_{k-1}\| \le M \|f(x_{k-1})\|.$$

Then $\|x_k - x_{k-1}\| \le \beta \|x_{k-1} - x_{k-2}\|$ (from induction hypo on $\|f(x_{k-1})\|$)

$$\Rightarrow \quad \|x_k - x_{k-1}\| \le \beta^{k-1} \|x_1 - x_0\|. \qquad \text{(by repeating)}$$

Thus, $\|x_k - x_0\| \le \sum_{j=1}^{k} \|x_j - x_{j-1}\| \le \left(\sum_{j=1}^{k} \beta^{j-1} \right) \|x_1 - x_0\|$

Thus, $\|x_k - x_0\| \leq \sum_{j=1}^{k} \|x_j - x_{j-1}\| \leq \left(\sum_{j=1}^{k} \beta^{j-1} \right) \|x_1 - x_0\|$

$$\leq \frac{\|x_1 - x_0\|}{1-\beta} \leq \frac{M}{1-\beta} \|f(x_0)\| \leq r$$

base case

by original assumption. (3)

which implies that $x_k \in B$.

Furthermore, since $x_k - x_{k-1} = -A_{k-1}^{-1}(x_\ell)(f(x_{k-1}))$,

$$f(x_{k-1}) = -A_{k-1}(x_\ell)(x_k - x_{k-1})$$

$$\Rightarrow f(x_k) = f(x_k) - f(x_{k-1}) - A_{k-1}(x_\ell)(x_k - x_{k-1}).$$

Again by the MVT,

$$\|f(x_k)\| \leq \sup_{x \in B} \|f'(x) - A_{k-1}(x_\ell)\| \, \|x_k - x_{k-1}\| \leq \frac{\beta}{M} \|x_k - x_{k-1}\|,$$

completing the induction. $\boxed{/}$

Lemma 2: f has a zero in B, and the convergence is geometric.

proof. $\|x_{k+j} - x_k\| \leq \sum_{i=0}^{j-1} \|x_{k+i+1} - x_{k+i}\| \leq \beta^k \left(\sum_{i=0}^{j-1} \beta^i \right) \|x_1 - x_0\| \leq \frac{\beta^k}{1-\beta} \|x_1 - x_0\|,$

for all $k \geq 0$ and $j \geq 0$, so (x_k) is a Cauchy sequence.

But B is a closed ball in a Banach space, so $\lim_{k \to \infty} x_k = a \in B$.

Since f is continuous on Ω (by differentiability),

$$\|f(a)\| = \lim_{k \to \infty} \|f(x_k)\| \leq \frac{\beta}{M} \lim_{k \to \infty} \|x_k - x_{k-1}\| = 0$$

$$\Rightarrow f(a) = 0.$$

Furthermore $\|x_{k+j} - x_k\| \leq \frac{\beta^k}{1-\beta} \|x_1 - x_0\|$

$$\Rightarrow \|x_k - a\| = \|a - x_k\| \leq \lim_{j \to \infty} \frac{\beta^k}{1-\beta} \|x_1 - x_0\| = \frac{\beta^k}{1-\beta} \|x_1 - x_0\|.$$

Lemma 3: f has a unique zero in B.

proof. Suppose $f(a) = f(b) = 0$, $a, b \in B$.

proof. Suppose $f(a) = f(b) = 0$, $a, b \in B$.

Note $A_0^{-1}(x_0)\big(A_0(x_0)\ (b-a)\big) = b-a$,

so $$b - a = -A_0^{-1}(x_0)\big(f(b) - f(a) - A_0(x_0)(b-a)\big)$$

$$\Rightarrow \|b-a\| \le \|A_0^{-1}(x_0)\| \sup_{x\in B} \|f'(x) - A_0(x_0)\| \, \|b-a\| \le \beta\|b-a\|.$$

Since $0<\beta<1$, $a=b$. $\boxed{\diagup}$

Although we've given a set of sufficient conditions, these are actually quite stringent.

Ex. Theorem 5.1 applies to $f(x) = x^2 - \alpha$, $\alpha > 0$ for any $x_0 > 0$ s.t.
$$\frac{6}{7}\alpha \le x_0^2 \le \frac{6}{5}\alpha,$$
with $\beta = \frac{2}{5}$, $r = \frac{1}{6}\cdot x_0$, $M = \frac{3}{5x_0}$. So x_0 has to be very close to $\sqrt{\alpha}$.

Note: Newton's method may converge even outside these conditions.

By being less general, we can relax the conditions a bit.

Thm 5.2/5.2 Let X be a Banach space and let $f: \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$. If $a \in \Omega$ is a pt s.t. $f(a) = 0$, if $f'(a)$ is a linear isomorphism, and if $\exists \lambda$, $0 < \lambda < \frac{1}{2}$ s.t.
$$\sup_{k\ge 0} \|A_k - f'(a)\|_{\mathcal{L}(X;Y)} \le \frac{\lambda}{\|(f'(a))^{-1}\|_{\mathcal{L}(Y;X)}},$$
then there is a closed ball B of center a s.t. $\forall x_0 \in B$, the sequence (x_k) given by
$$x_{k+1} = x_k - A_k^{-1}(f(x_k)), \quad k \ge 0,$$
is entirely contained within B and converges to a. Furthermore, a is the only 0 of f in B, and the convergence is geometric, i.e.
$$\|x_k - a\| \le \beta^k \|x_0 - a\|, \quad \text{for some } \beta < 1.$$

i.e. If we already know that we have a zero, then the isomorphisms A_k can be independent of x, assuming we start close to a.

$A_k(x) = f'(x)$ in the standard fashion, we can get

Or, if we use $A_k(x) = f'(x)$, in the standard fashion, we can get much stronger results.

Thm 5.3 (Newton–Kantorovich) See textbook for statement.

Now we can return to optimization, where we want zeros of the derivative $J' : \Omega \to E'$ of a function $J : \Omega \to \mathbb{R}$, $\Omega \subseteq E$.

Note J'' is a continuous bilinear form $J'' : E \times E \to \mathbb{R}$, which can be viewed as a linear map in $\mathcal{L}(E, E')$.
i.e. $J''(u)$ is a linear form given by $J''(u)(v) = J''(u, v)$.

Thm 5.4 Let E be a Banach space, let $J : \Omega \to \mathbb{R}$ be twice differentiable on the open subset $\Omega \subseteq E$, and assume $\exists$ constants r, M, β s.t. if

$$B = \{ x \in E \mid \|x - x_0\| \le r \} \subseteq \Omega,$$

then (1) $$\sup_{k \ge 0} \sup_{x \in B} \| A_k^{-1}(x) \|_{\mathcal{L}(E; E')} \le M$$

(2) $\beta < 1$ and $$\sup_{k \ge 0} \sup_{x \in B} \| J''(x) - A_k(x') \|_{\mathcal{L}(E; E')} \le \frac{\beta}{M}$$

(3) $$\| J'(x_0) \| \le \frac{r}{M}(1 - \beta)$$

Then the sequence (x_k) defined by

$$x_{k+1} = x_k - A_k^{-1}(x_\ell)(J'(x_k)), \qquad 0 \le \ell \le k$$

is entirely contained in B and converges to the unique zero a of J' geometrically.

Thm 5.5 Apply Thm 5.2 to J', and the $A_k(x)$ are isomorphisms in $\mathcal{L}(E, E')$ independent of $x \in \Omega$.

Note: When $E = \mathbb{R}^n$, Thm 5.4 gives iteration steps

$$x_{k+1} = x_k - A_k^{-1}(x_\ell) \nabla J(x_k), \qquad 0 \le \ell \le k,$$

where $\nabla J(x_k)$ is the gradient of J at x_k (identifying E' with $\mathbb{R}^n$).

For Newton's original method, $A_k = J''$, so

$$x_{k+1} = x_k - (\nabla^2 J(x_k))^{-1} \nabla J(x_k), \quad k \ge 0,$$

For Newton's original method, $A_k = \ldots$

$$x_{k+1} = x_k - \left(\nabla^2 J(x_k)\right)^{-1} \nabla J(x_k), \quad k \geq 0,$$

where $\nabla^2 J(x_k)$ is the Hessian of J at x_k.

Newton's method plays an important role in convex optimization, in particular for interior-pt methods. Also, variants of gradient descent can be viewed as generalized Newton's method.

Quadratic Optimization (Ch 6)

Let's consider in detail a couple common classes of quadratic optimization, before moving on to general results in optimization theory.

We will discuss minimizing $Q(x) = \frac{1}{2}x^TAx - x^Tb$ over
(1) $x \in \mathbb{R}^n$
(2) $x \in \mathbb{R}^n$ subject to linear or affine constraints
(3) x in the unit sphere.

This is important in practice because many energy functions can be defined in this form.

Def. 6.2 Given any symmetric $A \in \mathbb{R}^{n\times n}$, we write $A \succeq 0$ if A is pos. semi def.
$A \succ 0$ if A is pos. def.

Also, $A \succeq B$ if $A - B \succ 0$, a partial order on matrices called the pos. semidef. cone ordering.

Prop 6.2 Given a quadratic function,

$$Q(x) = \frac{1}{2} x^TAx - x^Tb,$$

if A is symmetric and $A \succ 0$, then $Q(x)$ has a unique global minimum at the solution of the linear system $Ax = b$. The minimum value of $Q(x)$ is

$$Q(A^{-1}b) = -\frac{1}{2} b^TA^{-1}b.$$

proof. Let $x = A^{-1}b$. Let $y \in \mathbb{R}^n$.

Then
$$\begin{aligned} Q(y) - Q(x) &= \tfrac{1}{2}y^TAy - y^Tb - \tfrac{1}{2}x^TAx + x^Tb \\ &= \tfrac{1}{2}y^TAy - y^TAx + \tfrac{1}{2}x^TAx \\ &= \tfrac{1}{2}(y-x)^TA(y-x) \geq 0 \end{aligned}$$

$$= \frac{1}{2}y^TAy - y^TAx + \frac{1}{2}x^TAx$$

$$= \frac{1}{2}(y-x)^TA(y-x) \geq 0$$

$$\Rightarrow Q(y) \geq Q(x).$$

$$\Rightarrow \min_{x \in \mathbb{R}} Q(x) = Q(A^{-1}b) = -\frac{1}{2}b^TA^{-1}b. \qquad \blacksquare$$

Aside: If $Q(x) = \frac{1}{2}x^TAx - x^Tb + c$, then $\operatorname{argmin}_{x \in \mathbb{R}^n} Q(x) = A^{-1}b$, but $Q(A^{-1}b) = -\frac{1}{2}b^TA^{-1}b + c$.

This allows us to recast a linear problem $Ax=b$ as a variational problem (finding the min. of an energy function). Often, we have additional constraints.

Def. 6.3 The quadratic constrained minimization problem consists in minimizing

$$Q(x) := \frac{1}{2}x^TA^{-1}x - b^Tx \quad \text{subject to linear constraints } B^Tx = f,$$

where $A^{-1} \in \mathbb{R}^{m \times m}$ is SPD, $B \in \mathbb{R}^{m \times n}$ has rank n, and where $b, x \in \mathbb{R}^m$ and $f \in \mathbb{R}^n$.

Note that we use A^{-1} instead of A because this constrained minimization has an interpretation as a set of equilibrium equations that give A. Notation taken from [Strang 1986].

The matrix $K = B^TAB$ is the stiffness matrix of e.g. a spring-mass system, or electrical networks, etc.

Recall that we can use Lagrange multipliers to solve this. The Lagrangian of the system is $L(x,\lambda) = Q(x) + \lambda^T(B^Tx - f) = \frac{1}{2}x^TA^{-1}x - (b - B\lambda)^Tx - \lambda^Tf$.

A necessary condition is $\nabla L(x,\lambda) = 0$.

$$\begin{cases} \dfrac{\partial L}{\partial x}(x,\lambda) = A^{-1}x - (b - B\lambda) = 0 \\[2ex] \dfrac{\partial L}{\partial \lambda}(x,\lambda) = B^Tx - f = 0 \end{cases}$$

$$\Rightarrow \left.\begin{aligned} A^{-1}x + B\lambda &= b \\ B^Tx &= f \end{aligned}\right\} \begin{pmatrix} A^{-1} & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}$$

$$\Rightarrow x = A(b - B\lambda)$$

$$\Rightarrow B^TA(b - B\lambda) = f$$

$$\Rightarrow B^TAB\lambda = B^TAb - f$$

$$\Rightarrow \lambda = (B^TAB)^{-1}(B^TAb - f), \qquad x = A(b - B\lambda).$$

$$\Rightarrow B^T A B \lambda = B^T A b - f$$
$$\Rightarrow \lambda = (B^T A B)^{-1}(B^T A b - f) \quad , \quad x = A(b - B\lambda).$$

(If we let $e = b - B\lambda$, we get equilibrium equations
$$\begin{cases} e = b - B\lambda \\ x = Ae \\ B^T x = f \end{cases} \qquad [\text{Strang, } 1986]$$
)

Let us define the dual function $G(\lambda)$ as follows:
$$G(\lambda) = \frac{1}{2}(B\lambda - b)^T A (B\lambda - b) + \lambda^T f.$$

Note that $\min_x L(x,\lambda) = L(A(b - B\lambda), \lambda)$ (by Prop 6.2)
$$= -G(\lambda).$$

Clearly, $L(x,\lambda) \geq -G(\lambda) \quad \forall x, \lambda$ because we minimized over x to get G.

But when $B^T x = f$, $L(x,\lambda) = Q(x)$, so
$$\forall \lambda, \quad \min_{x | B^T x = f} Q(x) = \min_{x | B^T x = f} L(x,\lambda) \geq \min_x L(x,\lambda) = -G(\lambda)$$

$$\Rightarrow \min_{x | B^T x = f} Q(x) \geq \max_\lambda -G(\lambda).$$

We are seeing here a special case of duality, which we will cover in more detail later.

Prop. 6.3 The quadratic constrained minimization problem has a unique solution (x,λ) given by
$$\begin{pmatrix} A^{-1} & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}$$

Furthermore, the component λ of the above solution is the unique value for which $-G(\lambda)$ is maximum.

proof. Let's compute $Q(x) + G(\lambda)$ subject to $B^T x = f$.
$$Q(x) + G(\lambda) = \frac{1}{2}x^T A^{-1} x - b^T x + \frac{1}{2}(B\lambda - b)^T A (B\lambda - b) + \lambda^T f$$
$$= \frac{1}{2}(A^{-1}x + B\lambda - b)^T A (A^{-1}x + B\lambda - b) \geq 0$$

and $Q(x) + G(\lambda) = 0$ iff $A^{-1}x + B\lambda - b = 0$
$$\Leftrightarrow A^{-1}x + B\lambda = b.$$

Then $Q(x) = -G(\lambda)$ exactly when $A^{-1}x + B\lambda = b$ (i.e. $Q(x) = \min_{x | B^T x = f} Q(x)$)

But $\min_{x | B^T x = f} Q(x) \geq \max_\lambda -G(\lambda)$ from above, so equality is achieved precisely at a constrained minimum of Q and an unconstrained maximum of $-G$. ☒

$x \mid B^T x = f \qquad \lambda$

precisely at a constrained minimum of Q and an unconstrained maximum of $-G$. $\square$