EECS 227C - Homework II

Walid Krichene

March 6, 2014

1 Let f be a convex, continuously differentiable function, with L-Lipschitz derivatives, and bounded level sets. That is, suppose that for any α, the set

$$\Omega_\alpha = \{x : f(x) \leq \alpha\}$$

is bounded. Prove that gradient decent with backtracking line search converges for all such f. Provide as tight an upper bound on the rate of convergence as you can.

The update for the gradient descent with backtracking line search is given by

$$x^{(k+1)} = x^{(k)} - t\nabla f(x_k)$$

where $t = \beta^i$ for the smallest i which satisfies the Armijo rule

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \alpha \left\langle -t\nabla f(x^k), \nabla f(x^{(k)}) \right\rangle$$

Using the fact that f has L-Lipschitz gradient, we have the following quadratic upper bound: for all x and y

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$$

applied to $x = x^{(k)}$ and $y = x^{(k+1)} = x^{(k)-t\nabla f(x^{(k)})}$, we have

$$\begin{aligned} f(x^{(k+1)}) &\leq f(x^{(k)}) - t\|\nabla f(x^{(k)})\|^2 + \frac{L}{2}t^2\|\nabla f(x^{(k)})\|^2 \\ &= f(x^{(k)}) - (t - \frac{L}{2}t^2)\|\nabla f(x^{(k)})\|^2 \end{aligned}$$

Let

$$h(t) = t - \frac{L}{2}t^2 - \alpha t = t(1 - \alpha - \frac{L}{2}t)$$

Thus

$$h(t) \geq 0 \Leftrightarrow t \in (0, \frac{2(1-\alpha)}{L})$$

In particular, the backtracking certainly stops if $t = \beta^i < \frac{2(1-\alpha)}{L}$, since in that case

$$f(x^{(k+1)}) \leq f(x^{(k)}) - (t - \frac{L}{2}t^2)\|\nabla f(x^{(k)})\|^2 \leq f(x^{(k)}) - \alpha t\|\nabla f(x^{(k)})\|^2$$

Therefore $t > \frac{2(1-\alpha)}{\beta L}$, and by the Armijo rule,

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{1}{\eta}\|\nabla f(x^{(k)})\|^2, \ \eta = \frac{\beta L}{2(1-\alpha)}$$

Summing these inequalities, we have

$$f(x^{(K+1)}) \leq f(x^{(0)}) - \frac{1}{\eta} \sum_{k=0}^{K} \|\nabla f(x^{(k)})\|^2$$

rearranging,

$$\sum_{k=0}^{K} \|\nabla f(x^{(k)})\|^2 \leq \eta(f(x^{(0)}) - f(x^{(K+1)})) \leq \eta(f(x^{(0)}) - f_{opt})$$

Thus

$$\min_{0 \leq k \leq K} \|\nabla f(x^{(k)})\| \leq \frac{\sqrt{\eta(f(x^{(0)}) - f_{opt})}}{\sqrt{K}}$$

Now by convexity we also have

$$f(x^{opt}) \geq f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x^{opt} - x^{(k)} \right\rangle$$

so by Cauchy-Schwartz

$$f(x^{(k)}) - f_{opt} \leq \|\nabla f(x^{(k)})\| \|x^{opt} - x^{(k)}\|$$

but since this is a descent method (guaranteed by the Armijo rule), $x^{(k)}$ remains in the level set $\Omega_{f(x^{(0)})}$, which is bounded by assumption. So if D is the diameter of $\Omega_{f(x^{(0)})}$, we have

$$0 \leq f(x^{(k)}) - f_{opt} \leq D\|\nabla f(x^{(k)})\|$$

and observing that $f(x^{(K)}) = \min_{0 \leq k \leq K} f(x^{(k)})$, we have

$$f(x^{(K)}) - f_{opt} \leq D \min_{0 \leq k \leq K} \|\nabla f(x^{(k)})\| \leq \frac{D\sqrt{\eta(f(x^{(0)}) - f_{opt}}}{\sqrt{N}}$$

and we have a $O(1/\sqrt{N})$ bound on the convergence rate.

2 Consider the momentum method

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}) + \beta_k (x^{(k)} - x^{(k-1)})$$

where α and β are chosen to minimize f at each iteration. Assume that $x_0 = x_{-1}$. Show that if $f(x) = \frac{1}{2}x^T Ax - b^T x$, then this method is equivalent to the conjugate gradient method.

Writing

$$p^{(k)} = \frac{1}{\alpha_k}(x_{k+1} - x_k)$$

We have

$$\begin{aligned} p^{(k)} &= -\nabla f(x^{(k)}) + \frac{\beta_k}{\alpha_k}(x^{(k)} - x^{(k-1)}) \\ &= -\nabla f(x^{(k)}) + \frac{\beta_k \alpha_{k-1}}{\alpha_k} p^{(k-1)} \end{aligned}$$

In the conjugate gradient method, we have

$$\begin{aligned} \alpha_k &= -\frac{\left\langle p^{(k)}, r^{(k)} \right\rangle}{\left\langle p^{(k)}, Ap^{(k)} \right\rangle} \\ \frac{\beta_k \alpha_{k-1}}{\alpha_k} &= \frac{\left\langle r^{(k)}, Ap^{(k-1)} \right\rangle}{\left\langle p^{(k-1)}, Ap^{(k-1)} \right\rangle} \end{aligned}$$

where $r^{(k)} = Ax^{(k)} - b = \nabla f(x^{(k)})$.

We change the optimization variables to α_k and $\gamma_k = \frac{\beta_k \alpha_{k-1}}{\alpha_k}$. Now for fixed $\gamma^{(k)}$, $p^{(k)}$ is fixed, and we can write

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \alpha_k p^{(k)}) \\ &= \frac{1}{2}\left\langle x^{(k)} + \alpha_k p^{(k)}, A(x^{(k)} + \alpha_k p^{(k)}) \right\rangle - \left\langle b, x^{(k)} + \alpha_k p^{(k)} \right\rangle \\ &= f(x^{(k)}) + \frac{1}{2}\alpha_k^2 \left\langle p^{(k)}, Ap^{(k)} \right\rangle + \alpha_k \left\langle Ax^{(k)}, p^{(k)} \right\rangle - \alpha_k \left\langle b, p^{(k)} \right\rangle \\ &= f(x^{(k)}) + \frac{1}{2}\alpha_k^2 \left\langle p^{(k)}, Ap^{(k)} \right\rangle + \alpha_k \left\langle r^{(k)}, p^{(k)} \right\rangle \end{aligned}$$

Setting the derivative to 0, we have

$$\alpha_k \left\langle p^{(k)}, Ap^{(k)} \right\rangle + \left\langle r^{(k)}, p^{(k)} \right\rangle = 0$$

thus

$$\alpha_k = -\frac{\left\langle p^{(k)}, r^{(k)} \right\rangle}{\left\langle p^{(k)}, Ap^{(k)} \right\rangle}$$

and we have the same expression for $\alpha^{(k)}$ as in the conjugate gradient method. Now plugging this value of α_k, we have

$$f(x^{(k+1)}) = f(x^{(k)}) - \frac{1}{2}\frac{\left\langle p^{(k)}, r^{(k)} \right\rangle^2}{\left\langle p^{(k)}, Ap^{(k)} \right\rangle}$$

now writing $p^{(k)} = -r^{(k)} + \gamma_k p^{(k-1)}$, we have

$$\begin{aligned} \frac{\left\langle p^{(k)}, r^{(k)} \right\rangle^2}{\left\langle p^{(k)}, Ap^{(k)} \right\rangle} &= \frac{\left\langle -r^{(k)} + \gamma_k p^{(k-1)}, r^{(k)} \right\rangle^2}{\left\langle -r^{(k)} + \gamma_k p^{(k-1)}, A(-r^{(k)} + \gamma_k p^{(k-1)}) \right\rangle} \\ &= \frac{(\gamma_k \left\langle p^{(k-1)}, r^{(k)} \right\rangle - \|r^{(k)}\|^2)^2}{\gamma_k^2 \left\langle p^{(k-1)}, Ap^{(k-1)}) \right\rangle - 2\gamma_k \left\langle p^{(k-1)}, Ar^{(k)} \right\rangle + \left\langle r^{(k)}, Ar^{(k)} \right\rangle} \end{aligned}$$

and the derivative w.r.t. γ_k is zero iff

$$\left\langle p^{(k-1)}, r^{(k)} \right\rangle \left(\gamma_k \left\langle p^{(k-1)}, r^{(k)} \right\rangle - \|r^{(k)}\|^2 \right) \left(\gamma_k^2 \left\langle p^{(k-1)}, Ap^{(k-1))} \right\rangle - 2\gamma_k \left\langle p^{(k-1)}, Ar^{(k)} \right\rangle + \left\langle r^{(k)}, Ar^{(k)} \right\rangle \right)$$
$$- \left(\gamma_k \left\langle p^{(k-1)}, Ap^{(k-1))} \right\rangle - \left\langle p^{(k-1)}, Ar^{(k)} \right\rangle \right) \left(\gamma_k \left\langle p^{(k-1)}, r^{(k)} \right\rangle - \|r^{(k)}\|^2 \right)^2 = 0$$

this is a second order polynomial in γ_k. I tried to show γ_k satisfies the conjugate gradient update, but it did not work.

Attempt 2: Writing $q^{(k)} = x^{(k+1)} - x^{(k)}$, and $\langle x, x \rangle_A = x^T A x$, we have

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} - \alpha_k r^{(k)} + \beta_k q^{(k)}) \\ &= \left\langle x^{(k)} - \alpha_k r^{(k)} + \beta_k q^{(k)}, x^{(k)} - \alpha_k r^{(k)} + \beta_k q^{(k)} \right\rangle_A - \left\langle x^{(k)} - \alpha_k r^{(k)} + \beta_k q^{(k)}, b \right\rangle \\ &= f(x^{(k)}) + \alpha_k^2 \left\langle r^{(k)}, r^{(k)} \right\rangle_A + \beta_k^2 \left\langle q^{(k)}, q^{(k)} \right\rangle_A + \\ &\beta_k \left\langle q^{(k)}, x^{(k)} \right\rangle_A - \alpha_k \left\langle r^{(k)}, x^{(k)} \right\rangle_A - 2\alpha_k \beta_k \left\langle q^{(k)}, r^{(k)} \right\rangle_A + \alpha_k \left\langle r^{(k)}, b \right\rangle - \beta_k \left\langle q^{(k)}, b \right\rangle \end{aligned}$$

writing the stationarity conditions, α_k, β_k are optimal only if

$$\begin{aligned} 2\alpha_k \left\langle r^{(k)}, r^{(k)} \right\rangle_A - \left\langle r^{(k)}, x^{(k)} \right\rangle_A - 2\beta_k \left\langle q^{(k)}, r^{(k)} \right\rangle_A + \left\langle r^{(k)}, b \right\rangle = 0 \\ 2\beta_k \left\langle q^{(k)}, q^{(k)} \right\rangle_A + \left\langle q^{(k)}, x^{(k)} \right\rangle_A - 2\alpha_k \left\langle q^{(k)}, r^{(k)} \right\rangle_A - \left\langle q^{(k)}, b \right\rangle = 0 \end{aligned}$$

i.e.

$$\begin{aligned} 2\alpha_k \left\langle r^{(k)}, r^{(k)} \right\rangle_A - \left\langle r^{(k)}, r^{(k)} \right\rangle - 2\beta_k \left\langle q^{(k)}, r^{(k)} \right\rangle_A = 0 \\ 2\beta_k \left\langle q^{(k)}, q^{(k)} \right\rangle_A + \left\langle q^{(k)}, r^{(k)} \right\rangle - 2\alpha_k \left\langle q^{(k)}, r^{(k)} \right\rangle_A = 0 \end{aligned}$$

i.e.

$$\beta_k = \frac{2\alpha_k \left\langle q^{(k)}, r^{(k)} \right\rangle_A - \left\langle q^{(k)}, r^{(k)} \right\rangle}{\left\langle q^{(k)}, q^{(k)} \right\rangle_A}$$

$$2\alpha_k \left\langle r^{(k)}, r^{(k)} \right\rangle_A - \left\langle r^{(k)}, r^{(k)} \right\rangle - 2 \left\langle q^{(k)}, r^{(k)} \right\rangle_A \frac{2\alpha_k \left\langle q^{(k)}, r^{(k)} \right\rangle_A - \left\langle q^{(k)}, r^{(k)} \right\rangle}{\left\langle q^{(k)}, q^{(k)} \right\rangle_A} = 0$$

3 Let ℓ and L be constant satisfying $0 < \ell \leq L$. Define

$$\alpha = \frac{4}{(\sqrt{L}+\sqrt{\ell})^2} \qquad\qquad \beta = \frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}$$

and the matrix

$$T(\lambda) = \begin{pmatrix} 1+\beta-\alpha\lambda & -\beta \\ 1 & 0 \end{pmatrix}$$

Assume throughout that $\lambda \in (\ell, L)$.

(a) Show that $T(\lambda)$ has two complex eigenvalues γ_1 and γ_2, each with magnitude β. The characteristic polynomial of $T(\lambda)$ is given by

$$\mathcal{X}_T(\lambda)(\gamma) = (\gamma - (1-\alpha\lambda+\beta))\gamma + \beta = \gamma^2 - (1-\alpha\lambda+\beta)\gamma + \beta$$

the discriminant is

$$\begin{aligned}\Delta &= (1-\alpha\lambda+\beta)^2 - 4\beta \\ &= \beta^2 - 2(1+\alpha\lambda)\beta + (1-\alpha\lambda)^2\end{aligned}$$

which is a second order polynomial in β. Its discriminant is $(1+\alpha\lambda)^2 - (1-\alpha\lambda)^2 = 4\alpha\lambda$

$$(1+\alpha\lambda) \pm 2\sqrt{\alpha\lambda} = (1 \pm \sqrt{\alpha\lambda})^2$$

therefore $\Delta \leq 0$ iff $(1-\sqrt{\alpha\lambda})^2 \leq \beta \leq (1+\sqrt{\alpha\lambda})^2$. The second inequality is satisfied since $\beta < 1$. For the first, we have $\ell \leq \lambda \leq L$, thus

$$\frac{2\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}} \leq \sqrt{\alpha\lambda} \leq \frac{2\sqrt{L}}{\sqrt{L}+\sqrt{\ell}}$$

or

$$\frac{\sqrt{\ell}-\sqrt{L}}{\sqrt{L}+\sqrt{\ell}} \leq 1-\sqrt{\alpha\lambda} \leq \frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}-\sqrt{\ell}}$$

thus

$$(1-\sqrt{\alpha\lambda})^2 \leq \beta^2 \leq \beta$$

which proves the inequality, and that $\Delta < 0$.

Therefore the roots of $\mathcal{X}_{T(\lambda)}$ are complex conjugate, given by

$$\gamma_{1,2} = \frac{1}{2}(1-\alpha\lambda+\beta) \pm i\frac{1}{2}\sqrt{4\beta - (1-\alpha\lambda+\beta)^2}$$

Thus

$$\|\gamma_1\|^2 = \|\gamma_2\|^2 = \frac{1}{4}(1-\alpha\lambda+\beta)^2 + \frac{1}{4}(4\beta - (1-\alpha\lambda+\beta)^2) = \beta$$

(b) Bound the condition number of the complex similarity transform $S(\lambda)$ which satisfies

$$S(\lambda)^{-1}T(\lambda)S(\lambda) = \begin{pmatrix} \gamma_1 & 0 \\ 0 & \bar{\gamma}_1 \end{pmatrix}$$

That is, compute $\max_\lambda \|S(\lambda)\|\|S(\lambda)^{-1}\|$.

First, we observe that $\begin{pmatrix} \gamma_1 \\ 1 \end{pmatrix}$ is an eigenvector of $T(\lambda)$ associated to the eigenvalue γ_1, since

$$\begin{pmatrix} 1+\beta-\alpha\lambda & -\beta \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \gamma_1 \\ 1 \end{pmatrix} = \begin{pmatrix} \gamma_1(1+\beta-\alpha\lambda)-\beta \\ \gamma_1 \end{pmatrix} = \begin{pmatrix} \gamma_1^2 \\ \gamma_1 \end{pmatrix} = \gamma_1\begin{pmatrix} \gamma_1 \\ 1 \end{pmatrix}$$

similarly, $\begin{pmatrix} \bar{\gamma}_1 \\ 1 \end{pmatrix}$ is an eigenvector for the eigenvalue $\bar{\gamma}_1$. Thus one possible similarity transform is given by

$$S(\lambda) = \begin{pmatrix} \gamma_1 & \bar{\gamma}_1 \\ 1 & 1 \end{pmatrix}$$

and

$$\kappa = \max_{\lambda} \|S(\lambda)\| \|S(\lambda)^{-1}\| = \frac{\sigma_{\max}(S(\lambda))}{\sigma_{\min}(S(\lambda))}$$

where the singular values of $S(\lambda)$ can be obtained by computing the eigenvalues of $S(\lambda)^* S(\lambda)$. Indeed, if the SVD of $S(\lambda)$ is given by $S(\lambda) = U^* \Sigma V$, then

$$S(\lambda)^* S(\lambda) = V^* \Sigma^2 V$$

so

$$\kappa^2 = \frac{\lambda_{\max}(S^* S)}{\lambda_{\min}(S^* S)}$$

We have

$$S(\lambda)^* S(\lambda) = \begin{pmatrix} \bar{\gamma}_1 & 1 \\ \gamma_1 & 1 \end{pmatrix} \begin{pmatrix} \gamma_1 & \bar{\gamma}_1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \beta + 1 & \bar{\gamma}_1^2 + 1 \\ \gamma_1^2 + 1 & \beta + 1 \end{pmatrix}$$

and its characteristic polynomial is

$$\mathcal{X}_{S^* S}(s) = (s - (\beta + 1))^2 - (\gamma_1^2 + 1)(\bar{\gamma}_1^2 + 1)$$

where

$$\begin{aligned} (\gamma_1^2 + 1)(\bar{\gamma}_1^2 + 1) &= \|\gamma_1\|^4 + \gamma_1^2 + \bar{\gamma}_1^2 + 1 \\ &= \beta^2 + 2\,\mathrm{Re}(\gamma_1^2) + 1 \\ &= \beta^2 + 2\frac{1}{4}((1 - \alpha\lambda + \beta)^2 + (1 - \alpha\lambda + \beta)^2 - 4\beta) + 1 \\ &= \beta^2 + (1 - \alpha\lambda + \beta)^2 - 2\beta + 1 \end{aligned}$$

so

$$\mathcal{X}_{S^* S}(s) = s^2 - 2(\beta + 1)s + 4\beta - (1 - \alpha\lambda + \beta)^2$$

with discriminant

$$(\beta + 1)^2 - 4\beta + (1 - \alpha\lambda + \beta)^2 = (\beta - 1)^2 + (1 - \alpha\lambda + \beta)^2 \geq 0$$

thus the eigenvalues are

$$\beta + 1 \pm \sqrt{(\beta - 1)^2 + (1 - \alpha\lambda + \beta)^2}$$

thus

$$\begin{aligned} \kappa^2 &= \max_{\ell \leq \lambda \leq L} \frac{\beta + 1 + \sqrt{(\beta - 1)^2 + (1 - \alpha\lambda + \beta)^2}}{\beta + 1 - \sqrt{(\beta - 1)^2 + (1 - \alpha\lambda + \beta)^2}} \\ &= \max_{\ell \leq \lambda \leq L} -1 + 2\frac{\beta + 1}{\beta + 1 - \sqrt{(\beta - 1)^2 + (1 - \alpha\lambda + \beta)^2}} \end{aligned}$$

which is maximal when $h(\lambda) = (1 - \alpha\lambda + \beta)^2$ is maximal. Since h is quadratic, it is maximal at $\lambda = \ell$ or $\lambda = L$. We have

$$h(\ell) = (1 - \alpha\lambda + \beta)^2 = (1 - (\frac{2\sqrt{\ell}}{\sqrt{\ell} + \sqrt{L}})^2 + \beta)^2 = (1 - (1 - \beta)^2 + \beta)^2 = (-\beta^2 + 3\beta)^2$$

$$h(L) = (1 - \alpha\lambda + \beta)^2 = (1 - (\frac{2\sqrt{L}}{\sqrt{\ell} + \sqrt{L}})^2 + \beta)^2 = (1 - (1 + \beta)^2 + \beta)^2 = (\beta^2 + \beta)^2$$

but $-\beta^2+3\beta \geq \beta^2+\beta$ (because $0 \leq \beta \leq 1$) thus the maximum is attained at $\lambda = \ell$, and

$$\kappa^2 = -1 + 2\frac{\beta+1}{\beta+1-\sqrt{(\beta-1)^2+(3\beta-\beta^2)^2}}$$

this is what the function looks like

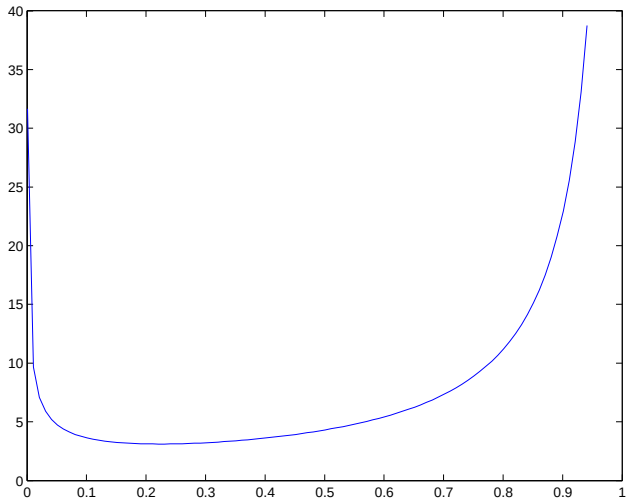

Figure 1: $\kappa(\beta)$ for $\beta \in (0,1)$

4 Recall the Heavy Ball method, we tried to force the matrix

$$T = \begin{pmatrix} (1+\beta)I - \alpha A & -\beta I \\ I & 0 \end{pmatrix}$$

to have as small a spectral as possible when the eigenvalues of A were in the interval (ℓ, L). What is the smallest *norm* of T possible? That is, compute

$$\Gamma = \min_{\alpha>0, 0<\beta<1} \max_{\ell I \preceq A \preceq L I} \|T\|$$

What values of α and β achieve the minimum?

By letting $U = (1+\beta)I - \alpha A$, we have

$$\ell I \preceq A \preceq L I \Leftrightarrow (1+\beta-\alpha L)I \preceq U \preceq (1+\beta-\alpha\ell)I$$

thus

$$\min_{\alpha>0, 0<\beta<1} \max_{(1+\beta-\alpha L)I \preceq U \preceq (1+\beta-\alpha\ell)I} \| \begin{pmatrix} U & -\beta I \\ I & 0 \end{pmatrix} \|$$

The operator norm of a matrix M is given by

$$\|M\|^2 = \max_x \frac{\|Mx\|^2}{\|x\|^2} = \max_x \frac{x^T M^T M x}{\|x\|^2}$$

here

$$M^T M = \begin{pmatrix} U^T & I \\ -\beta I & 0 \end{pmatrix} \begin{pmatrix} U & -\beta I \\ I & 0 \end{pmatrix} = \begin{pmatrix} U^T U + I & -\beta U^T \\ -\beta U & \beta^2 I \end{pmatrix}$$

and writing $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, we have

$$x^T M^T M x = \|Ux_1\|^2 + \|x_1\|^2 - 2\beta (Ux_1)^T x_2 + \beta^2 \|x_2\|^2$$

thus the Reighley quotient becomes

$$\frac{\|Ux_1\|^2 + \|x_1\|_2 - 2\beta (Ux_1)^T x_2 + \beta^2 \|x_2\|^2}{\|x_1\|^2 + \|x_2\|^2}$$

we observe that for any given x_1, x_2, we can always increase the quotient by rotating x_2 so that $x_2 \propto -Ux_1$, so $-2\beta (Ux_1)^T x_2 = 2\beta \|Ux_1\| \|x_2\|$. Then we want to maximize

$$\frac{\|Ux_1\|^2 + \|x_1\|^2 + 2\beta \|Ux_1\| \|x_2\| + \beta^2 \|x_2\|^2}{\|x_1\|^2 + \|x_2\|^2} = \frac{1 + (\frac{\|Ux_1\|}{\|x_1\|} + \beta \frac{\|x_2\|}{\|x_1\|})^2}{1 + \frac{\|x_2\|}{\|x_1\|}^2}$$

Now we have

$$(1+\beta-\alpha L)I \preceq U \preceq (1+\beta-\alpha\ell)I \Leftrightarrow \lambda(\alpha,\beta) \leq \frac{\|Ux_1\|}{\|x_1\|} \leq \Lambda(\alpha,\beta)$$

where

$$\Lambda(\alpha,\beta) = \max(|1+\beta-\alpha\ell|, |1+\beta-\alpha L|)$$
$$\lambda(\alpha,\beta) \text{ is the projection of 0 on the interval } (1+\beta-\alpha\ell|, 1+\beta-\alpha L|)$$

And with the change of variable $u = \frac{\|Ux_1\|}{\|x_1\|}$, $r = \frac{\|x_2\|}{\|x_1\|}$, the problem becomes

$$\Gamma^2 = \min_{\alpha,\beta} \max_{\substack{r>0 \\ \lambda(\alpha,\beta) \leq u \leq \Lambda(\alpha,\beta)}} \frac{1 + (u+\beta r)^2}{1+r^2}$$

For any fixed r, β, $u \mapsto \frac{1+(u+\beta r)^2}{1+r^2}$ is increasing, thus the maximum is attained at $u = \Lambda(\alpha, \beta) = \max(|1 + \beta - \alpha\ell|, |1 + \beta - \alpha L|)$.

Therefore the problem becomes

$$\Gamma^2 = \min_{\alpha,\beta} \max_{r>0} \frac{1 + (\Lambda(\alpha, \beta) + \beta r)^2}{1 + r^2}$$

Let $h(r, \alpha, \beta) = \frac{1+(\Lambda(\alpha,\beta)+\beta r)^2}{1+r^2}$. For all β, r, h is minimal when α minimizes $\Lambda(\alpha, \beta)$. From Figure 2, $\Lambda(\alpha, \beta) = \max(|1+\beta-\alpha L|, |1+\beta-\alpha\ell|)$ is minimal when the two functions are equal, i.e. when $-(1+\beta-\alpha L) = 1+\beta-\alpha\ell$. That is,

$$\alpha^*(\beta) = \frac{2(1 + \beta)}{L + \ell}$$

then $\Lambda(\alpha^*(\beta), \beta) =$. Finally,

$$\Gamma^2 = \min_{0<\beta<1} \max_{r>0} \frac{1 + (\frac{(1+\beta)(L-\ell)}{L+\ell} + \beta r)^2}{1 + r^2}$$

and the function is increasing in β for all r, therefore $\beta^* = 0$, and

$$\Gamma^2 = \max_{r>0} \frac{1 + (\frac{L-\ell}{L+\ell})^2}{1 + r^2} = 1 + (\frac{L - \ell}{L + \ell})^2$$

attained at $\beta^* = 0$, $\alpha^* = \frac{2}{L+\ell}$

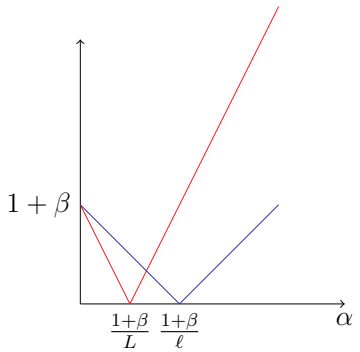

Figure 2: $|1 + \beta - \alpha L|$ and $|1 + \beta - \alpha\ell|$ as a function of α

5 Let f be convex, continuously differentiable function, with L-Lipschitz derivatives. Prove that Nesterov's method converges when $\alpha_k = \frac{1}{L}$ and $\beta_k = \frac{3}{k+2}$ for all such f, Provide as tight an upper bound on the rate of convergence as you can.

Using the same notation as in class, we have the following update:

$$y_k = x_k + \beta_k(x_k - x_{k-1})$$
$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$$

Starting the analysis in the same way, we have by induction

$$\begin{aligned} f(x_k) - f_* \leq & \prod_{i=0}^{k}(1-\theta_i)(f(x(0)) - f_*) \\ & + \frac{L}{2}\prod_{i=1}^{k}(1-\theta_i)\|(1-\theta_0)x_0 + \theta_0 x^* - y_0\|^2 \\ & + \frac{L}{2}\sum_{i=1}^{k}\left[\prod_{j=i+1}^{k}(1-\theta_j)\right]\left[\|(1-\theta_i)x_i + \theta_i x^* - y_i\|^2 - (1-\theta_i)\|(1-\theta_{i-1})x_{i-1} + \theta_{i-1}x^* - x_i\|^2\right] \\ & - \frac{L}{2}\|(1-\theta_{k-1})x_{k-1} + \theta_{k-1}x^* - x_k\|^2 \end{aligned}$$

We keep the first and second lines, ignore the last line (negative term), and choose the sequence θ_k and β_k so that the term on the third line is negative. A sufficient condition is that

$$(1-\theta_i)x_i + \theta_i x^* - y_i = \alpha_i(1-\theta_i)^{\frac{1}{2}}((1-\theta_{i-1})x_{i-1} + \theta_{i-1}x^* - x_i) \tag{1}$$

for some $\alpha_i \in (-1, 1)$. If this is the case, then $\|(1-\theta_i)x_i + \theta_i x^* - y_i\| < (1-\theta_i)^{\frac{1}{2}}\|(1-\theta_{i-1})x_{i-1} + \theta_{i-1}x^* - x_i\|$ and we have the desired inequality.

Pluggin the expression of $y_i = x_i + \beta_i(x_i - x_{i-1})$ in (1), we have

$$(1-\theta_i)x_i + \theta_i x^* - x_i - \beta_i x_i + \beta_i x_{i-1} = \alpha(1-\theta_i)^{\frac{1}{2}}((1-\theta_{i-1})x_{i-1} + \theta_{i-1}x^* - x_i)$$

and identifying terms, it suffices to have

$$\begin{aligned} \theta_i &= \alpha_i(1-\theta_i)^{\frac{1}{2}}\theta_{i-1} \\ (1-\theta_i) - 1 - \beta_i &= -\alpha_i(1-\theta_i)^{\frac{1}{2}} \\ \beta_i &= \alpha_i(1-\theta_i)^{\frac{1}{2}}(1-\theta_{i-1}) \end{aligned}$$

i.e.

$$\frac{\theta_i}{\theta_{i-1}} = \alpha_i(1-\theta_i)^{\frac{1}{2}} \tag{2}$$
$$\beta_i = -\theta_i + \frac{\theta_i}{\theta_{i-1}} \qquad \text{using (2)}$$
$$\beta_i = \frac{\theta_i}{\theta_{i-1}}(1-\theta_{i-1}) \qquad \text{using (2)}$$

the two conditions on β_i are the same. Now, choosing $\alpha_i = (1-\theta_i)$, we have $\alpha_i \in (0,1)$ (since $\theta_i \in (0,1)$), and the recursive relation on θ_i becomes

$$\frac{\theta_i}{\theta_{i-1}} = (1-\theta_i) \tag{3}$$

This is verified for example by $\theta_i = \frac{1}{i+2}$, since

$$\frac{\theta_i}{\theta_{i-1}} = \frac{1/(i+2)}{1/(i+1)} = \frac{i+1}{i+2} = 1 - \frac{1}{i+2} = 1 - \theta_i$$

With this expression for θ_i, we have

$$\beta_i = \frac{\theta_i}{\theta_{i-1}}(1 - \theta_{i-1}) = \frac{i+1}{i+2}\frac{i}{i+1} = \frac{i}{i+2}$$

Continuing with the analysis, we have

$$\begin{aligned} f(x_k) - f_* &\leq \left[\prod_{i=1}^{k}(1-\theta_i)\right]\left[(1-\theta_0)(f(x_0) - f_*) + \|(1-\theta_0)x_0 + \theta_0 x^* - y_0\|^2\right] \\ &= C_0\left[\prod_{i=1}^{k}(1-\theta_i)\right] \end{aligned}$$

where C_0 is a constant which only depends on the initial point x_0. Now using (3), we have

$$\prod_{i=1}^{k}(1-\theta_i) = \prod_{i=1}^{k}\frac{\theta_i}{\theta_{i-1}} = \frac{\theta_k}{\theta_0} = \frac{2}{k+2}$$

Therefore we obtain the bound

$$f(x_k) - f_* \leq \frac{2}{k+2}C_0$$

where $C_0 = (1-\theta_0)(f(x_0) - f_*) + \|(1-\theta_0)x_0 + \theta_0 x^* - y_0\|^2$.

6 Apply Newton's method with a constant stepsize to minimize the function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = \frac{1}{3}\left(\sum_{i=1}^{n} x_i^2\right)^{\frac{3}{2}}$$

Identify the range of stepsizes for which this method converges. Show that for any stepsize within this range, the iterates converge linearly to $x_{\text{opt}} = 0$. Explain why the method does not converge quadratically to the optimal solution.

The (damped) update for the Newton method is given by

$$x^{(k+1)} = x^{(k)} - t(\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)})$$

where t is the step size. We have

$$\frac{\partial f}{\partial x_j}(x) = x_j\left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$$

$$\frac{\partial^2 f}{\partial x_j x_k}(x) = \begin{cases} \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}} + x_j^2\left(\sum_{i=1}^{n} x_i^2\right)^{-\frac{1}{2}} & \text{if } k = j \\ x_j x_k\left(\sum_{i=1}^{n} x_i^2\right)^{-\frac{1}{2}} & \text{otherwise} \end{cases}$$

therefore

$$\nabla f(x) = \|x\|x$$

$$\nabla^2 f(x) = \frac{1}{\|x\|}xx^T + \|x\|I$$

we observe that x is an eigenvector of $\nabla^2 f(x)$ since

$$\nabla^2 f(x)x = (\frac{1}{\|x\|}xx^T + \|x\|I)x = \frac{1}{\|x\|}x\|x\|^2 + \|x\|x = 2\|x\|x$$

therefore

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

i.e.

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

Finally, the Newton update is given by

$$x^{(k+1)} = x^{(k)} - \frac{t}{2}x^{(k)} = (1 - \frac{t}{2})x^{(k)}$$

and by induction

$$x^{(k)} = (1 - \frac{t}{2})^k x^{(0)}$$

which converges to 0 (linearly) iff $t \in (0, 2)$.

We obtain linear convergence, and not quadratic. The theorem of quadratic convergence does not apply here because $\lim_{x\to 0} \nabla^2 f(x) = 0$, and one of the assumptions in the theorem is that $\nabla^2 f(0) \succ \ell I$ for some positive ℓ.

7 Derive closed form expression for the following quasi-Newton methods. In both cases, determine whether $H^{(k)}$ is positive definite. The updates are given by

$$x^{(k+1)} = x^{(k)} - H^{(k)}\nabla f(x^{(k)})$$

The the secant condition is

$$s^{(k)} = H^{(k+1)}y^{(k)}$$

where $s^{(k)}$ and $y^{(k)}$ are determined at the end of iterate $k+1$:

$$s^{(k)} = x^{(k+1)} - x^{(k)} = -H^{(k)}\nabla f(x^{(k)})$$
$$y^{(k)} = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$$

(a) $H^{(k+1)} = \text{diag}(h_1, \ldots, h_n)$ matrix for all k. Minimize the distance to the last iterate in the operator norm.

Let $H^{(k+1)} = \text{diag}(h_1^{(k+1)}, \ldots, h_n^{(k+1)})$. Then $\|H^{(k+1)} - H^{(k)}\| = \max_i |h_i^{(k+1)} - h_i^{(k)}|$, and the problem is

$$\min_{\{h_i^{(k+1)}\}} \max_i |h_i^{(k+1)} - h_i^{(k)}|$$
$$\text{s.t.} h_i^{(k+1)} y_i^{(k)} = s_i^{(k)}$$

and the solution is

$$h_i^{(k+1)} = \begin{cases} \frac{s_i^{(k)}}{y_i^{(k)}} & \text{if } y_i^{(k)} \neq 0 \\ h_i^{(k)} & \text{if } y_i^{(k)} = s_i^{(k)} = 0 \end{cases}$$

$H^{(k+1)}$ is positive definite if $s_i^{(k)}$ and $y_i^{(k)}$ have the same sign.

(b) $H^{(k+1)}$ a rank-one update $H^{(k+1)} = H^{(k)} + vv^T$.

By definition of the update, $H^{(k)} \succ 0 \Rightarrow H^{(k+1)} \succ 0$.

In this case the constraint can be written as

$$(H^{(k)} + vv^T)y^{(k)} = s^{(k)}$$

that is

$$vv^T y^{(k)} = s^{(k)} - H^{(k)}y^{(k)}$$

Define $b^{(k)} = s^{(k)} - H^{(k)}y^{(k)}$. Then the equation is $vv^T y^{(k)} = b^{(k)}$, and if v is a solution, then $v = \alpha b^{(k)}$, where α satisfies

$$\alpha^2 b^{(k)}(b^{(k)})^T y^{(k)} = b^{(k)}$$

that is

$$\alpha^2 (b^{(k)})^T y^{(k)} = 1$$

Therefore we have a solution only if $\left\langle b^{(k)}, y^{(k)} \right\rangle > 0$, in which case

$$v = \frac{1}{\sqrt{\left\langle b^{(k)}, y^{(k)} \right\rangle}} b^{(k)}$$

The update is feasible only if $\left\langle b^{(k)}, y^{(k)} \right\rangle > 0$, i.e.

$$\left\langle s^{(k)} - H^{(k)}y^{(k)}, y^{(k)} \right\rangle > 0$$

that is,

$$\left\langle s^{(k)}, y^{(k)} \right\rangle > \left\langle y^{(k)}, H^{(k)} y^{(k)} \right\rangle$$

which is a stronger condition than the one we had in BFGS, where we required ($\left\langle s^{(k)}, y^{(k)} \right\rangle > 0$).

A sufficient condition in the strong convex case: by definition, $\left\langle y^{(k)}, s^{(k)} \right\rangle = \left\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle$, and if f is ℓ-strongly convex with L-Lipschitz gradient, we have

$$\left\langle s^{(k)}, y^{(k)} \right\rangle \geq \frac{\ell}{2} \left\langle s^{(k)}, s^{(k)} \right\rangle$$
$$\left\langle y^{(k)}, y^{(k)} \right\rangle \leq L^2 \left\langle s^{(k)}, s^{(k)} \right\rangle$$

thus

$$\left\langle s^{(k)}, y^{(k)} \right\rangle \geq \frac{\ell}{2L^2} \left\langle y^{(k)}, y^{(k)} \right\rangle$$

so if $\frac{\ell}{2L^2} > \lambda_{\max}(H^{(k)})$, we have

$$\left\langle s^{(k)}, y^{(k)} \right\rangle \geq \frac{\ell}{2L^2} \left\langle y^{(k)}, y^{(k)} \right\rangle > \lambda_{\max}(H^{(k)}) \left\langle y^{(k)}, y^{(k)} \right\rangle \geq \left\langle y^{(k)}, H^{(k)} y^{(k)} \right\rangle$$

so the condition is satisfied and $H^{(k+1)}$ can be constructed.