Unconstrained Optimization II

Note Title 4/1/2022

To understand the convergence properties of descent methods, we begin with a case study:

- The objective is a strictly convex quadratic, i.e. $f(x) = c + b^T x + \frac{1}{2} x^T A x$, $A \succ 0$
- The descent direction is $-\nabla f(x)$ (i.e. the GD method)
- The step size selection: constant, exact, backtracking

Note 1: If a general $f: \mathcal{U} \subseteq \mathbb{R}^n \to \mathbb{R}$ has a strict local minimizer at x^* with $\nabla^2 f(x^*) \succ 0$, then for $x \approx x^*$,

$$f(x) \approx f(x^*) + \cancel{\nabla f(x^*)^T (x - x^*)}^{\,0} + \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*),$$ which is a strictly convex quadratic.

Note 2: But x^* being a strict local minimizer of f does not imply $\nabla^2 f(x^*) \succ 0$.

E.g. $f(x) = x^4$ has $x^* = 0$ as the unique global minimizer, but $f''(0) = 0$.

As we shall see, the convergence properties of GD does depend on the positive definiteness of $\nabla^2 f(x^*)$.

Back to the case study, $f(x) = c + b^T x + \frac{1}{2} x^T A x$, $A \succ 0$, has its unique global minimizer at

$$x^* = -A^{-1} b.$$

$$f(x) = \frac{1}{2}(x - x^*) A (x - x^*) + (\text{constant}).$$

check this invariance

By translation invarance of GD (with either constant or exact line search), we may assume WLOG that

$$f(x) = \frac{1}{2} x^T A x. \qquad (\nabla f(x) = Ax)$$

Constant step size: $x_{k+1} = x_k - \bar{\alpha} A x_k = (I - \bar{\alpha} A) x_k$ How to choose $\bar{\alpha}$?

Exact line search: $\alpha_k = \operatorname{argmin}_\alpha f(x_k - \alpha d_k)$ write $d_k = \nabla f(x_k) = Ax_k$

$$= \operatorname{argmin}_\alpha \frac{1}{2}(x_k - \alpha d_k)^T A (x_k - \alpha d_k)$$

$$= \operatorname{argmin}_\alpha \frac{1}{2}\alpha^2 d_k^T A d_k - \alpha \underbrace{d_k^T A x_k}_{= d_k} + \frac{1}{2} x_k^T A x_k$$

$$= \underbrace{d_k^T d_k / d_k^T A d_k}$$

$\downarrow$

So $x_{k+1} = x_k - \alpha_k A x_k$

[What tools do we have for analyzing discrete dynamical system besides eigen-analysis ?]

In the constant step size case ,

$$\lim_{k\to\infty} x_k = 0 \text{ , for any } x_0 \iff \text{Spectrum of } I-\bar{\alpha}A \subset (-1,1).$$

$$\text{Spectrum of } I-\bar{\alpha}A = \{1-\bar{\alpha}\lambda_i : i=1,\cdots,n\}$$

$0<\lambda_1\leq\cdots\leq\lambda_n$ are the eigenvalues of A

$$|1-\bar{\alpha}\lambda_i|<1 \iff 0<\bar{\alpha}<2/\lambda_i \ \forall i \iff 0<\bar{\alpha}<2/\lambda_n$$

So convergence is guaranteed if the constant step is chosen so that $\bar{\alpha}\in(0,2/\lambda_n)$.

What about rate of convergence ?

$$x_k = (I-\bar{\alpha}A)^k x_0 \text{ , } \quad A = U[\lambda_1 \cdots \lambda_n]U^T \text{ , } \quad I-\bar{\alpha}A = U\,\mathrm{diag}(1-\bar{\alpha}\lambda_1,\cdots,1-\bar{\alpha}\lambda_n)U^T$$

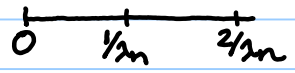
$$\|x_k\|_2 = \|U\,\mathrm{diag}(1-\bar{\alpha}\lambda_1,\cdots,1-\bar{\alpha}\lambda_n)^k U^T x_0\|_2$$

$$\leq \underbrace{\left(\max_i |1-\bar{\alpha}\lambda_i|\right)^k}_{=\rho(I-\bar{\alpha}A)} \|x_0\|_2$$

ρ: spectral radius

(This rate can be attained with an appropriate choice of x_0.)

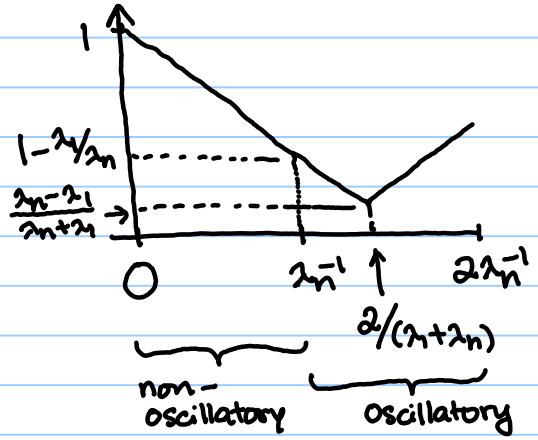
Solving $\min_{\bar{\alpha}} \rho(I - \bar{\alpha} A)$



Since $0 < \lambda_1 \leq \cdots \leq \lambda_n$

$$1 - \bar{\alpha}\lambda_1 \geqslant \cdots \geqslant 1 - \bar{\alpha}\lambda_n \qquad \bar{\alpha} \in (0, 1/\lambda_n] \iff 1 - \bar{\alpha}\lambda_i \geqslant 0, \ \forall i$$

$$(*) \quad \max_{i=1,\cdots,n} |1 - \bar{\alpha}\lambda_i| = \begin{cases} 1 - \bar{\alpha}\lambda_1 & \text{when } \bar{\alpha} \in (0, 2/(\lambda_1+\lambda_n)] \\ -(1 - \bar{\alpha}\lambda_n) & \text{when } \bar{\alpha} \in [2/(\lambda_1+\lambda_n), 2/\lambda_n) \end{cases}$$



So the optimal step size is

$$\bar{\alpha} = 2/(\lambda_1 + \lambda_n), \quad \text{with optimal}$$

(worst case) rate of convergence $O\left(\left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^k\right)$.

↖ worst among all initial vectors x_0.

Ex: Prove (*).

- For any choice of step size $\bar{\alpha} \in (0, 2\lambda_n^{-1})$. We have

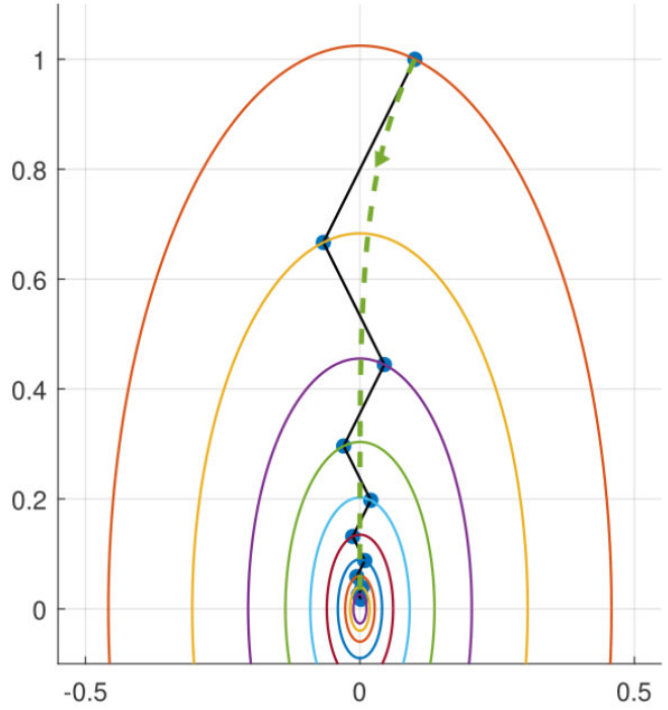
$$\|x_k\|_2 \leq \rho(I - \bar{\alpha}A)^k \|x_0\|.$$

The ROC on the RHS is tight as long as x_0 has a non-zero component of the eigenvector associated with the spectral radius of $I - \bar{\alpha}A$.

So the "worst case" ROC is actually generic, and is attained by almost all initial vectors x_0.

- The optimal ROC worsens as $\frac{\lambda_n}{\lambda_1} \to \infty$, since $\rho = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \to 1$ as $\lambda_n/\lambda_1 \to \infty$.

- One needs to know λ_1, λ_n in order to get the optimal step size $2/(\lambda_1 + \lambda_n)$. (But computing $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ is not much easier than computing $A^{-1}b$!)



- nothing oscillatory for GF
- x_R oscillatory when $\bar{\alpha} \in [1/\lambda_n, 2/\lambda_n)$

Eg. $A = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$, $x_0 = \begin{bmatrix} 10^{-1} \\ 1 \end{bmatrix}$

$\bar{\alpha} = 2/(5+1) = 1/3$

Green curve: solution of the gradient flow

$$x'(t) = -Ax(t), \quad x(0) = x_0 \Rightarrow x(t) = \begin{bmatrix} 10^{-1} e^{-5t} \\ e^{-t} \end{bmatrix}$$

GD can be thought of a time-discretization of GF.

But, unlike in numerical ODE, the goal is not to solve the ODE accurately. To the very least, we do not want step size $= \Delta t \to 0$.

What about exact line search?

$$x_{k+1} = x_k - \alpha_k A x_k, \quad \alpha_k = d_k^T d_k / d_k^T A d_k, \quad d_k = A x_k$$

$$= (I - \alpha_k A) x_k$$

↑ a linear map, but varies with k (in a somewhat nonlinear way)

To say the most obvious, eigen-analysis 101 does not apply anymore!

But the linearity still helps:

$$f(x_{k+1}) = x_{k+1}^T A x_{k+1} = (x_k - \alpha_k d_k)^T A (x_k - \alpha_k d_k) = x_k^T A x_k - 2\alpha_k \underbrace{d_k^T A x_k}_{d_k} + \alpha_k^2 d_k^T A d_k$$

$$= x_k^T A x_k - 2 \frac{d_k^T d_k}{d_k^T A d_k} d_k^T d_k + \left(\frac{d_k^T d_k}{d_k^T A d_k}\right)^2 d_k^T A d_k$$

$$= x_k^T A x_k - \frac{(d_k^T d_k)^2}{d_k^T A d_k}$$

$$= \underbrace{x_k^T A x_k}_{f(x_k)} \left[1 - \frac{(d_k^T d_k)^2}{(d_k^T A d_k)\underbrace{(x_k^T A x_k)}_{d_k^T A^{-1} A A^{-1} d_k}}\right] = \left[1 - \frac{(d_k^T d_k)^2}{(d_k^T A d_k)(d_k^T A^{-1} d_k)}\right] f(x_k)$$

↑ ✡

Note $f(x_k) = \frac{1}{2} x_k^T A x_k$,
$= \frac{1}{2} \|x_k\|_A^2$, where $\|x\|_A := \sqrt{x^T A x}$ is a norm.

By norm equivalence, $c_1 \|\cdot\|_A \le \|\cdot\| \le c_2 \|\cdot\|_A$, the ROC of $\|x_k\|$ to 0 in any norm is the same as that of $\|x_k\|_A$ $(= \sqrt{2 f(x_k)})$.

So if we can obtain an upper bound of ✡, say θ^2, then at least we can conclude that

$$\|x_k\| = O(\theta^k)$$

But we have seen the expression in ✡ before!

By Kantorovich inequality,

$$✡ \le 1 - 4\lambda_1\lambda_n/(\lambda_1+\lambda_n)^2 = (\lambda_n-\lambda_1)^2/(\lambda_n+\lambda_1)^2$$

so $\|x_k\|_A \le \left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^k \|x_0\|_A$ (no hidden constant if we use the A-norm)

Note: This is the same ROC we got with the optimal constant step size, achieved without needing to first compute λ_1 and λ_n!

Ex: Kantorovich vs Rayleigh

In estimating $\frac{(x^Tx)^2}{(x^TAx)(x^TA^{-1}x)}$, we may bypass the tricky Kantorovich inequality and use only the basic result of Rayleigh quotient:

$$\frac{(x^Tx)^2}{(x^TAx)(x^TA^{-1}x)} = \frac{1}{(x^TAx/x^Tx)(x^TA^{-1}x/x^Tx)} \geq \frac{1}{\lambda_{max}(A)\,\lambda_{max}(A^{-1})} = \frac{1}{\lambda_n \lambda_1^{-1}} = \frac{\lambda_1}{\lambda_n} = R$$

Q: compared to Kantorovich's bound $4\lambda_1\lambda_n/(\lambda_1+\lambda_n)^2 = K$, which bound is better? by how much?

Hint: analyze K/R, notice that this ratio depends only on $\kappa(A) = \lambda_n/\lambda_1$.

What happens when A is very well-conditioned? very ill-conditioned?

Compare:

Optimal constant step size

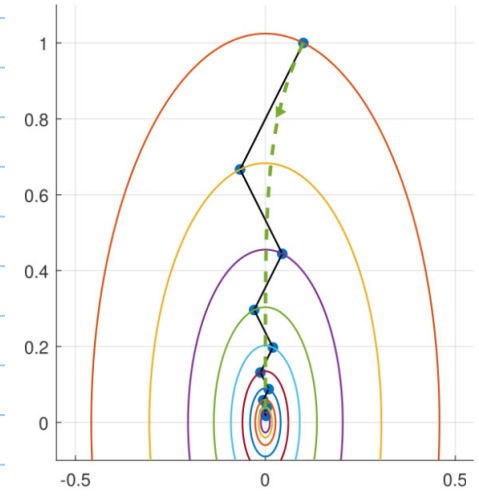
$$\alpha = 2/(\lambda_1 + \lambda_n)$$

Exact line search

$$\alpha_k = d_k^T d_k / d_k^T A d_k \quad , \quad d_k = A x_k = \nabla f(x_k)$$

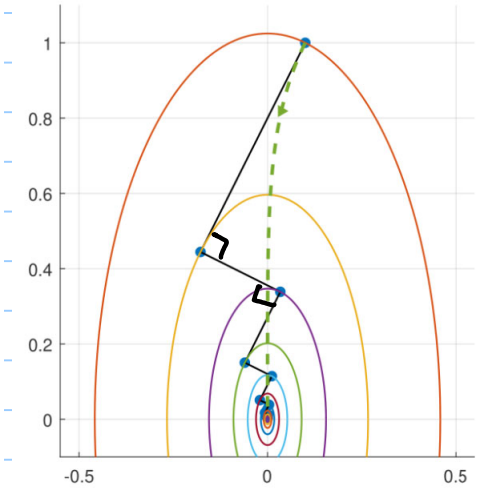
↑ no need to know λ_1, λ_n,
(exploits the fact that f is a quadratic)

$$A = \begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}$$

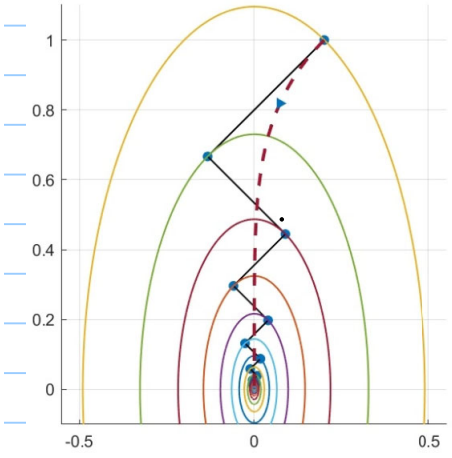


Optimal constant

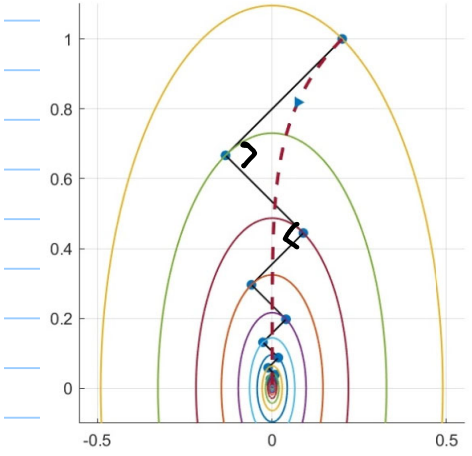
$$x_0 = \begin{bmatrix} 1/10 \\ 1 \end{bmatrix}$$



exact – converge faster than "optimal constant"



Optimal constant



exact

$$x_0 = \begin{bmatrix} 1/5 \\ 1 \end{bmatrix}$$

While the worst case ROC for the two methods are the same ($O\left(\left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^k\right)$), and for the optimal step size method this upper bound is tight for generic x_0, it is actually not tight for most initial vectors x_0 when A has only two distinct eigenvalues, but it is quite tight for most initial vectors x_0 when A has at least three distinct eigenvalues and λ_1/λ_n is large. (see ROC plot on next page.)

Not surprisingly, the analysis of the precise ROC of the exact line search method, and its dependence on the initial guess x_0 is very tricky.

A specific fact about exact line search, borrowed from [Beck]:

Lemma 4.7. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient method with exact line search for solving a problem of minimizing a continuously differentiable function f. Then for any $k=0,1,2,\ldots$*

↑ not just quadratics

$$(\mathbf{x}_{k+2}-\mathbf{x}_{k+1})^T(\mathbf{x}_{k+1}-\mathbf{x}_k)=0.$$

Proof. By the definition of the gradient method we have that $\mathbf{x}_{k+1}-\mathbf{x}_k=-t_k\nabla f(\mathbf{x}_k)$ and $\mathbf{x}_{k+2}-\mathbf{x}_{k+1}=-t_{k+1}\nabla f(\mathbf{x}_{k+1})$. Therefore, we wish to prove that $\nabla f(\mathbf{x}_k)^T\nabla f(\mathbf{x}_{k+1})=0$. Since

$$t_k\in \operatorname{argmin}_{t\geq 0}\{g(t)\equiv f(\mathbf{x}_k-t\nabla f(\mathbf{x}_k))\},$$

and the optimal solution is not $t_k=0$, it follows that $g'(t_k)=0$. Hence,

$$-\nabla f(\mathbf{x}_k)^T\nabla f(\mathbf{x}_k-t_k\nabla f(\mathbf{x}_k))=0,$$

meaning that the desired result $\nabla f(\mathbf{x}_k)^T\nabla f(\mathbf{x}_{k+1})=0$ holds. □

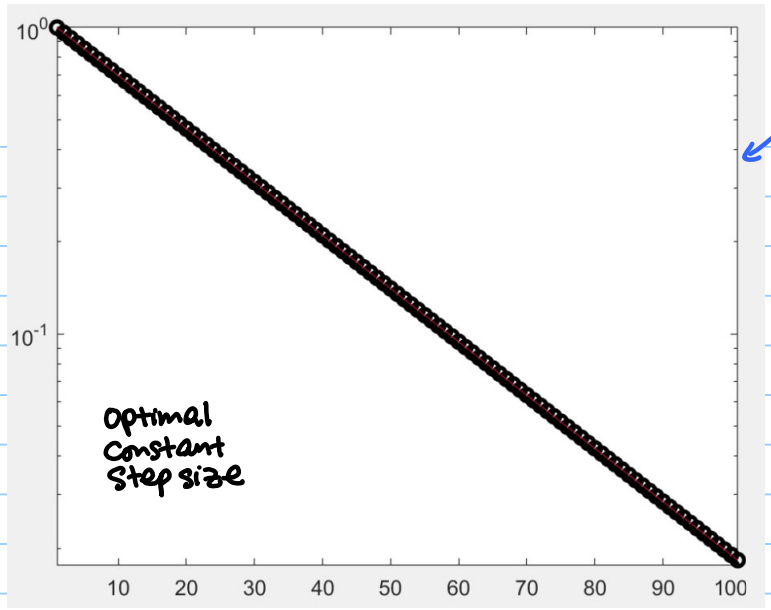
ROC plots

$n=2$

$A=\begin{bmatrix} 50 & \\ & 1 \end{bmatrix}$

1000 trials

x_0 uniformly sampled from $\{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$



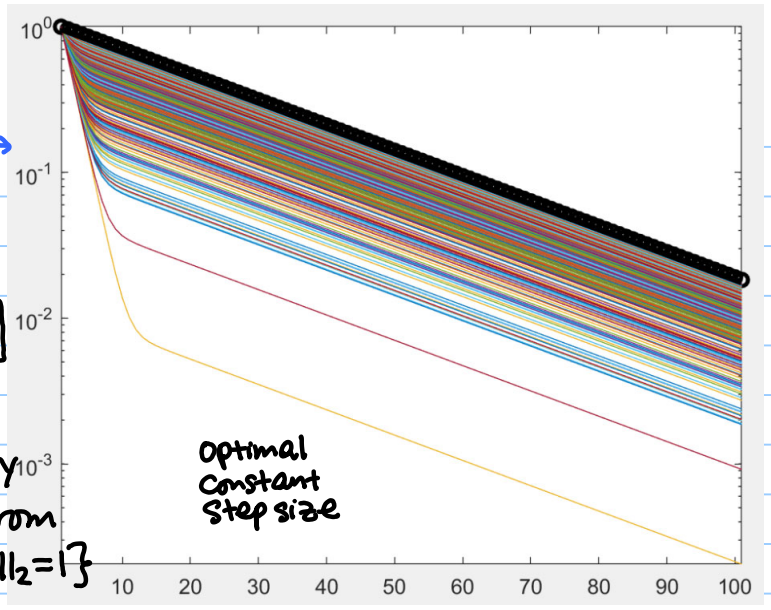
"Lin. Alg. 101"

$n=3$

$A=\begin{bmatrix} 50 & & \\ & 10 & \\ & & 1 \end{bmatrix}$

1000 trials

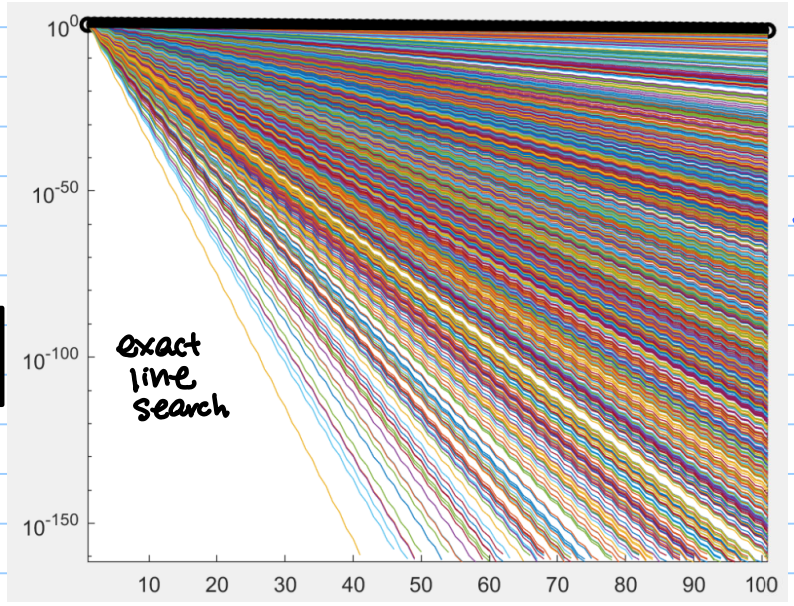
x_0 uniformly sampled from $\{x \in \mathbb{R}^3 : \|x\|_2 = 1\}$



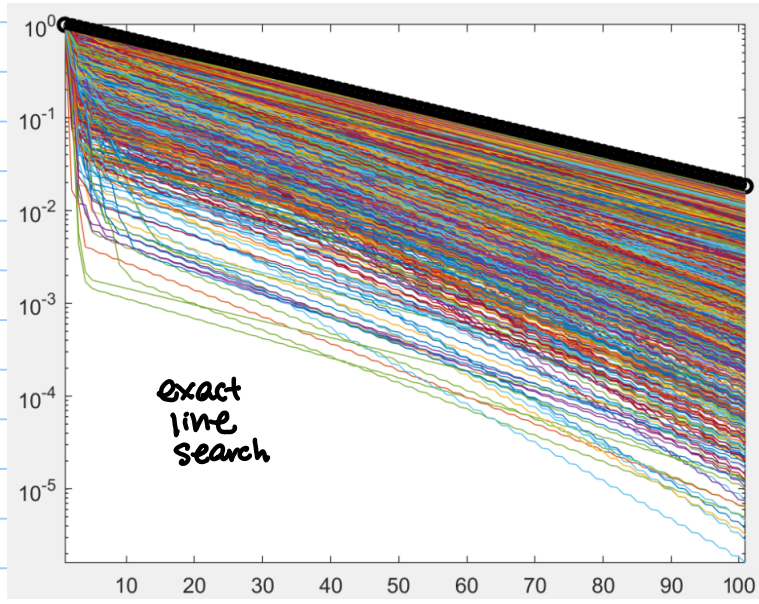
Black line in all 4 plots:

k vs $\log \rho^k$

$$\rho = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}$$



harder to analyze



I will send you my code for generating the four plots.

Note that when $e_k \sim C\theta^k$, then $\log e_k \sim \log C + k \log \theta$

This means a plot of k vs $\log e_k$, or a semilogy( ) (in Matlab) plot of k vs e_k, looks like a straightline with slope $\log \theta$.

Ex: modify my code to explore what happens to backtracking

see the following pages for an analysis of backtracking applied to quadratics.

What about backtracking?

Recall: The optimal constant step size rule $\bar{\alpha} = 2/(\lambda_1+\lambda_n)$,

A suboptimal constant step size selection: $\alpha \in (0, 2\lambda_n^{-1})$

Exact linesearch step size: $\alpha_k = \nabla f_k^T \nabla f_k / \nabla f_k^T A \nabla f_k$, $\nabla f_k = Ax_k + b$

} all require some global information of the objective, which may or may not be available in practice for general objective functions.

Backtracking is convenient in practice as its computation only requires being able to compute $f(x)$ and $\nabla f(x)$.

Recall: In backtracking (applied to GD) we seek a small enough step size α_k for which

$$f(x_k) - f(x_k - \alpha_k \underbrace{\nabla f(x_k)}_{-p_k}) \geq c\,\alpha_k \underbrace{\|\nabla f(x_k)\|^2}_{-\nabla f_k^T p_k} \quad \text{— (SD)}$$

Either

(i) $\alpha_k = \bar{\alpha}$ (initial value of the step size chosen in the backtracking algorithm)

OR

(ii) α_k/ρ is not acceptable and does not satisfy (SD), i.e.

$$f(x_k) - f(x_k - \tfrac{\alpha_k}{\rho}\nabla f(x_k)) < c\,\tfrac{\alpha_k}{\rho}\|\nabla f(x_k)\|^2 \quad \text{— (1)}$$

Notice the following "sufficient decrease property" of GD applied to $f(x) = \frac{1}{2}x^TAx + b^Tx$:

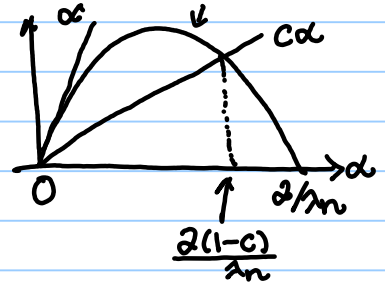
$$f(x) - f(x - \alpha\nabla f(x)) = \tfrac{1}{2}x^TAx + b^Tx - \tfrac{1}{2}[x - \alpha\nabla f(x)]^TA[x - \alpha\nabla f(x)] - b^T[x - \alpha\nabla f(x)]$$

$$= \alpha\nabla f(x)^TAx - \tfrac{1}{2}\alpha^2\nabla f(x)^TA\nabla f(x) + \alpha\nabla f(x)^Tb$$

$$= \alpha\nabla f(x)^T\nabla f(x) - \tfrac{1}{2}\alpha^2\nabla f(x)^TA\nabla f(x)$$

$$\nabla f(x)^TA\nabla f(x) \le \lambda_n \nabla f(x)^T\nabla f(x)$$

$$f(x) - f(x - \alpha\nabla f(x)) \ge \alpha\left(1 - \frac{\alpha\lambda_n}{2}\right)\|\nabla f(x)\|_2^2 \quad \leftarrow ②$$



② ⇒ (SD) is satisfied as long as $\alpha_k \in [0, 2(1-c)/\lambda_n]$

① ⇒ α_k/ρ does not satisfy (SD)

So ① + ② ⇒ $\alpha_k/\rho > 2(1-c)/\lambda_n$,

i.e. $\alpha_k > \frac{2(1-c)\rho}{\lambda_n}$. ← this (uniform in k) lower bound of step size is attributable to both the backtracking algorithm and that $A = \nabla^2 f(x) \preceq \lambda_n I$.

(i) + (ii) $\Rightarrow \alpha_k \geq \min\{\bar{\alpha}, \frac{2(1-c)\rho}{\lambda_n}\}$.

$$f(x_k) - f(x_{k+1}) \geq c\alpha_k \|\nabla f(x_k)\|_2^2 \geq c \min\{\bar{\alpha}, \tfrac{2(1-c)\rho}{\lambda_n}\} \underbrace{\|\nabla f(x_k)\|_2^2}_{\geq \lambda_1 2f(x_k) \text{ (see below)}} \quad \text{——③}$$

Note: $$\frac{\nabla f_k^T A^{-1} \nabla f_k}{\nabla f_k^T \nabla f_k} \leq \lambda_{\max}(A^{-1}) = \lambda_1^{-1}$$

so $$\|\nabla f_k\|_2^2 \geq \lambda_1 (Ax_k)^T A^{-1} (Ax_k) = \lambda_1 x_k^T A x_k = \lambda_1 2f(x_k)$$

By ③:

$$f(x_{k+1}) \leq f(x_k)\left[1 - 2\lambda_1 c \min\{\bar{\alpha}, \tfrac{2(1-c)\rho}{\lambda_n}\}\right]^{\theta}$$

$\theta = 1 - \min\{2\lambda_1 c\bar{\alpha}, \underbrace{4c(1-c)\rho\frac{\lambda_1}{\lambda_n}}_{\leq 1}\} < 1$. And $\theta \to 1$ as $\frac{\lambda_1}{\lambda_n} \to 0$.

<1

<1

Later, we shall see that the results above can be generalized to any general convex objective function f satisfying $\alpha_1 I \preceq \nabla^2 f(x) \preceq \lambda_n I$.

But let's first see how to establish convergence without assuming $\alpha_1 I \preceq \nabla^2 f(x)$

Note that in this case we cannot expect linear convergence (i.e. $O(\rho^k)$, $\rho<1$).

E.g. $f(x)=x^\beta$, $f''(x)=\beta(\beta-1)x^{\beta-2}$, $0 \le f''(x) \le \underbrace{\beta(\beta-1)}_{=:L}$ $\quad x\in[0,1]$

$f''(0)=0$ for $\beta>2$.

GD: $\quad x_{n+1} = x_n - \frac{1}{L}\nabla f(x_n) = x_n - \frac{1}{\beta-1}x_n^{\beta-1}$

Note that $x_0 \in (0,1) \Rightarrow x_n > 0 \ \forall n$ and $x_n \downarrow$ (not clear if $x_n \downarrow 0$ as this point)

log-log plots of n vs $f(x_n)$

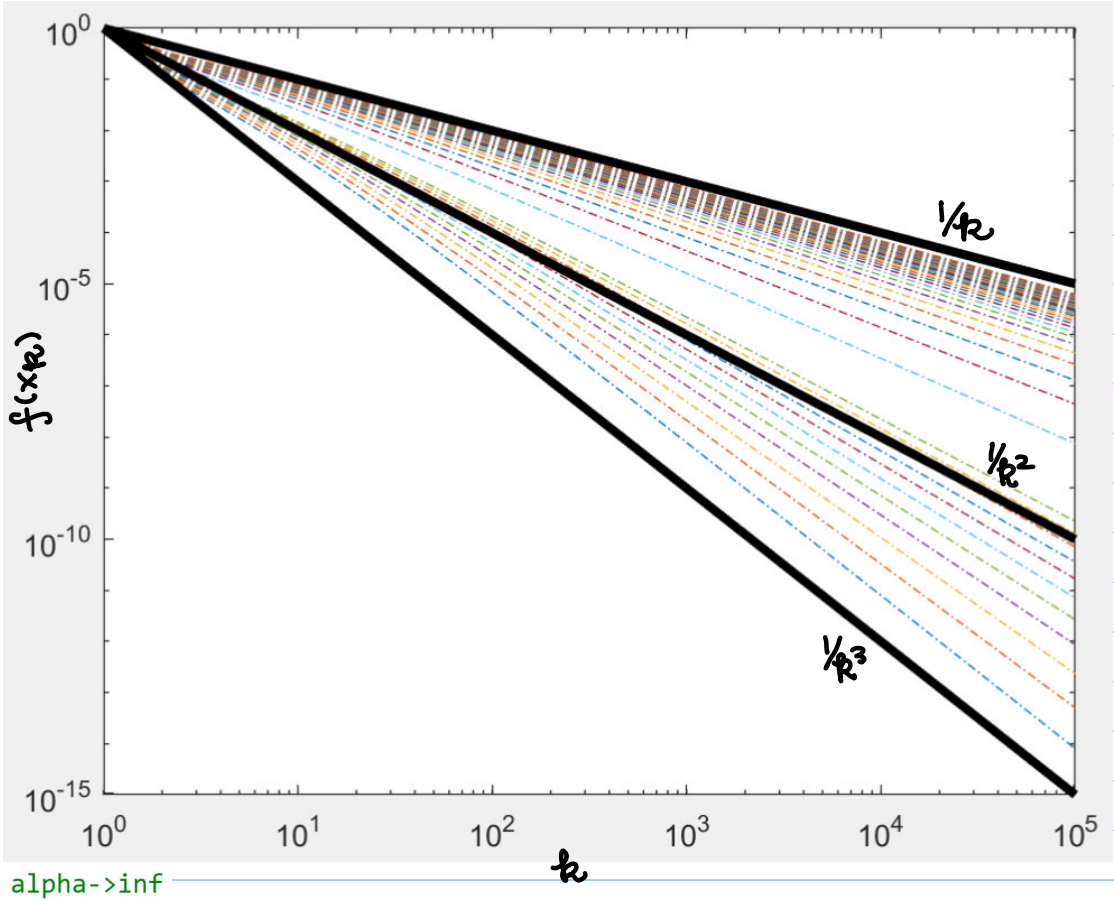
if $f(x_n) \sim C \cdot 1/n^\alpha$ then $\log f(x_n) \sim \log(C) - \alpha \log n$

straight line with slope $-\alpha$.

Computational results:

```
% GD applied to non-strongly convex objectives
N = 100000;
% x^beta
for beta = 3:.1:4
    x=zeros(1,N+1);
    x(1) = 1;
    for k=1:N
        x(k+1) = x(k)-1/(beta*(beta-1))*beta*x(k)^(beta-1);
    end
    F = x.^beta;
    loglog(1:(N+1),F,'-.') % 1/k^2 convergence
    hold on
end

for beta = 4:50
    x=zeros(1,N+1);
    x(1) = 1;
    for k=1:N
        x(k+1) = x(k)-1/(beta*(beta-1))*beta*x(k)^(beta-1);
    end
    F = x.^beta;
    loglog(1:(N+1),F,'-.') % 1/k^2 for alpha=4, approach 1/k as alpha->inf
    hold on
end
loglog(1:(N+1), (1:(N+1)).^(-1), 'k', 'LineWidth', 3)
loglog(1:(N+1), (1:(N+1)).^(-2), 'k', 'LineWidth', 3)
loglog(1:(N+1), (1:(N+1)).^(-3), 'k', 'LineWidth', 3)
```



Analysis

$$x_{n+1} = x_n - \frac{1}{\beta-1} x_n^{\beta-1} = x_n\left(1 - \frac{1}{\beta-1} x_n^{\beta-2}\right) \qquad (\beta > 2)$$

$$x_{n+1}^{-\alpha} = x_n^{-\alpha}\left(1 - \frac{1}{\beta-1} x_n^{\beta-2}\right)^{-\alpha} \qquad (1-y)^{-\alpha} \approx 1 + \alpha y, \; y \approx 0$$

$$\approx x_n^{-\alpha}\left(1 + \frac{\alpha}{\beta-1} x_n^{\beta-2}\right)$$

$$x_{n+1}^{-(\beta-2)} \approx x_n^{-(\beta-2)} + \frac{\beta-2}{\beta-1} \quad \text{if } \alpha = \beta-2$$

So $x_n^{-(\beta-2)} \approx x_0^{-(\beta-2)} + \frac{\beta-2}{\beta-1} n$

$$x_n \approx \left[x_0^{-(\beta-2)} + \frac{\beta-2}{\beta-1} n\right]^{-\frac{1}{\beta-2}} \qquad f(x_n) \approx \left[x_0^{-(\beta-2)} + \frac{\beta-2}{\beta-1} n\right]^{-\beta/\beta-2}$$

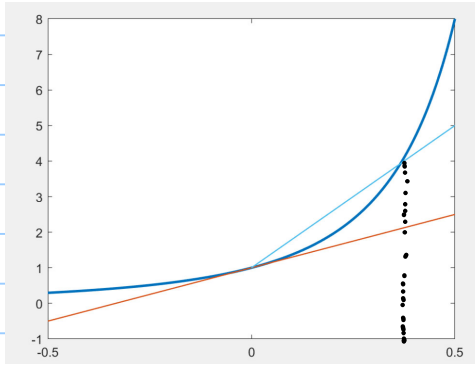
$$= \left(\frac{\beta-1}{\beta-2}\right)^{\frac{1}{\beta-2}} n^{-\frac{1}{\beta-2}} (1 + o(1)) \qquad \boxed{f(x_n) \approx \left(\frac{\beta-1}{\beta-2}\right)^{\frac{\beta}{\beta-2}} n^{-\beta/\beta-2} (1 + o(1))}$$

It is easy to make this analysis rigorous, based on

$$1 + \alpha y \leq (1-y)^{-\alpha} \leq 1 + (\alpha + \varepsilon) y$$

↑ holds for all $y < 1$ by convexity

↑ holds for $y \in [0, \delta(\varepsilon)]$



Fix $\varepsilon > 0$, $\quad x_{n+1}^{-(\beta-2)} = x_n^{-(\beta-2)}\left(1 - \frac{1}{\beta-1}x_n^{\beta-2}\right)^{-(\beta-2)} \le x_n^{-(\beta-2)}\left(1 + \frac{\beta-2+\varepsilon}{\beta-1}x_n^{\beta-2}\right)$ for large enough n, say $n \ge n_0(\varepsilon)$

$$= x_n^{-(\beta-2)} + \frac{\beta-2+\varepsilon}{\beta-1}$$

So $\quad x_n^{-(\beta-2)} \le x_{n_0}^{-(\beta-2)} + \frac{\beta-2+\varepsilon}{\beta-1}(n-n_0)$

$$x_n \ge \left[c'n + c\right]^{-\frac{1}{\beta-2}} \qquad (c > 0)$$

$$f(x_n) \ge \left[c'n + c\right]^{-\frac{\beta}{\beta-2}} \qquad n \ge n_0 \qquad -(LB)$$

This means x_n and $f(x_n)$ converge to 0 sublinearly (i.e. slower than $O(\rho^n)$ for any $\rho < 1$.)

In proving the lower bound (LB) above, we assumed (in the first line) that $x_n \to 0$. To see that it is indeed true, note that:

$$x_{n+1}^{-(\beta-2)} = x_n^{-(\beta-2)}\left(1 - \frac{1}{\beta-1}x_n^{\beta-2}\right)^{-(\beta-2)} \ge x_n^{-(\beta-2)}\left(1 + \frac{\beta-2}{\beta-1}x_n^{\beta-2}\right)$$

$$= x_n^{-(\beta-2)} + \beta-2/\beta-1.$$

So $\quad x_n^{-(\beta-2)} \ge x_0^{-(\beta-2)} + n\frac{\beta-2}{\beta-1}$

$$x_n \le \left[C''n + C\right]^{-\beta/\beta-2} \quad \forall n \ge 0.$$