New Directions 2016: Mathematical Optimization
August 2016
Worksheet 2

1. Suppose $Y_t \in \mathbb{R}^{n \times d}$ has orthonormal columns and let

$$-\nabla f(Y) = -(I - YY^T)\frac{df(Y)}{dY}\bigg|_{Y=Y_t} =: U\Sigma V^T$$

be the thin SVD of the negative gradient, so $U \in \mathbb{R}^{n \times d}$, $\Sigma \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{d \times d}$. Show that the Grassmannian update

$$Y_{t+1} = Y_t V \cos(\Sigma \eta_t) V^T + U \sin(\Sigma \eta_t) V^T$$

results in another matrix with orthonormal columns.

2. Find positive values of $\bar{\epsilon}$, γ_1, and γ_2 such that the Gauss-Southwell choice $d^k = -[\nabla f(x^k)]_{i_k} e_{i_k}$, where e_{i_k} is the unit vector $(0, \dots, 0, 1, 0, \dots, 0)^T$ with the 1 in position i_k. where $i_k = \arg\max_{i=1,2,\dots,n} |[\nabla f(x^k)]_i|$ satisfies conditions

$$-(d^k)^T \nabla f(x^k) \geq \bar{\epsilon} \|\nabla f(x^k)\| \|d^k\|, \quad \gamma_1 \|\nabla f(x^k)\| \leq \|d^k\| \leq \gamma_2 \|\nabla f(x^k)\|.$$

3. Consider a line-search method for min $f(x)$ in which the search direction d_k satisfies the conditions

$$-d_k^T \nabla f(x_k) \geq \bar{\epsilon} \|\nabla f(x_k)\| \|d_k\|, \quad \|d_k\| \geq \gamma_1 \|\nabla f(x_k)\|.$$

for positive $\bar{\epsilon}$ and γ_1. The steps have the form

$$x_{k+1} = x_k + \alpha_k d_k, \quad \text{for some } \alpha_k > 0.$$

Suppose that we use a backtracking procedure to select α_k, where we try in turn $\alpha_k = \bar{\alpha}, \bar{\alpha}/2, \bar{\alpha}/4, \dots,$ for some $\bar{\alpha} > 0$, stopping when the following sufficient decrease condition is satisfied:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k d_k^T \nabla f(x_k),$$

for some constant $c_1 \in (0, 1)$. Prove that this backtracking procedure yields the following reduction in f at each iteration:

$$f(x_{k+1}) \leq f(x_k) - \Delta \|\nabla f(x_k)\|^2,$$

for some $\Delta > 0$, and find an appropriate value for Δ. Hints: (a) consider separately the cases of $\alpha_k = \bar{\alpha}$ and $\alpha_k < \bar{\alpha}$ (that is, whether backtracking was needed, or not); (b) Note that when backtracking is required, the *previous* value of α_k tried (namely, $2\alpha_k$) must have failed the sufficient decrease test.

4. Show that Nesterov's optimal method applied to the convex quadratic $f(x) = \frac{1}{2}x^TQx - b^Tx + c$ (where Q is a symmetric positive definite matrix whose eigenvalues lie in the range $[m, L]$ for $0 < m < L$) yields a linear convergence rate that is approximately the same as for the heavy ball method. The analysis should follow closely the analysis of the heavy-ball method shown in the notes. Proceed in the following steps.

   (a) Note that Nesterov's optimal method is given by the formula

   $$x^{k+1} = x^k - \alpha\nabla f(x^k + \beta(x^k - x^{k-1})) + \beta(x^k - x^{k-1}).$$

   Specialize this formula to the particular case of f convex quadratic, and find a matrix T such that

   $$w^k = Tw^{k-1}, \quad k = 1, 2, \ldots,$$

   where

   $$w^k := \begin{bmatrix} x^{k+1} - x^* \\ x^k - x^* \end{bmatrix}.$$

   (b) Following the technique used in the heavy-ball analysis, show that by a similarity transformation, we can transform T to a matrix

   $$\begin{bmatrix} T_1 & 0 & \ldots & 0 \\ 0 & T_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & T_n \end{bmatrix},$$

   where each T_i is a 2×2 block that depends on the ith eigenvalue of Q (that we denote by λ_i). Write out the form of T_i.

   (c) Find the eigenvalues of each T_i, as a function of α, β, and λ_i.

   (d) Show that for the choices

   $$\alpha = 1/L, \quad \beta = \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}},$$

   we these eigenvalues are all bounded in magnitude by $1 - \sqrt{m/L}$.

5. Minimize a quadratic objective $f(x) = (1/2)x^T Ax$ with some first-order methods, generating the problems using the following code fragment to generate a Hessian with eigenvalues in the range $[m, L]$.

```
mu=0.01; L=1; kappa=L/mu;
n=100;
A = randn(n,n); [Q,R]=qr(A);
D=rand(n,1); D=10.^D; Dmin=min(D); Dmax=max(D);
D=(D-Dmin)/(Dmax-Dmin);
D = mu + D*(L-mu);
A = Q'*diag(D)*Q;
epsilon=1.e-6;
kmax=1000;
x0 = randn(n,1); % use a different x0 for each of the 10 trials
```

Run the code in each case until $f(x_k) \leq \epsilon$ for tolerance $\epsilon = 10^{-6}$. Implement the following methods.

- Steepest descent with $\alpha_k \equiv 2/(m + L)$.
- Steepest descent with exact line search.
- Heavy-ball method, with $\alpha = 4/(\sqrt{L} + \sqrt{m})^2$ and $\beta = (\sqrt{L} - \sqrt{m})/(\sqrt{L} + \sqrt{m})$.
- Nesterov's optimal method, with $\alpha = 1/L$ and $\beta = (\sqrt{L} - \sqrt{m})/(\sqrt{L} + \sqrt{m})$.

(a) Tabulate the average number of iterations required, over 10 random starts.

(b) Draw a plot of the convergence behavior on a typical run, plotting iteration number against $\log_{10}(f(x_k) - f(x^*))$. (Use the same figure, with four different colors for the four algorithms.)

(c) Discuss your results, noting in particular whether the worst-case convergence analysis is reflected in the practical results.

6. Discuss happens to the codes and algorithms in the previous question when we reset m to 0 (making f weakly convex).

7. Prove that Moreau's prox-operator is a contraction.

8. Show that a closed proper convex function h and its Moreau envelope $M_{\lambda,h}$ have identical minimizers.

9. Calculate $\text{prox}_{\lambda h}(x)$ and $M_{\lambda,h}(x)$ for $h(x) = \frac{1}{2}\|x\|_2^2$.