10-725 Convex Optimization
Homework 1

Shashank Singh*

Due September 19, 2013

*sss1@andrew.cmu.edu

1 Mastery set [25 points] (Aaditya)

A1 [2] $\forall k \in 1, \ldots, n$, define $S_k := \sum_{i=1}^k \theta_i$ and $y_k := \displaystyle\sum_{i=1}^k \frac{\theta_i x_i}{S_k} \in C$. Suppose that, for some $k \in 1, \ldots, n-1$, $y_k \in C$. Then,

$$\begin{aligned}
y_{k+1} := \sum_{i=1}^{k+1} \frac{\theta_i x_i}{S_{k+1}} &= \frac{\theta_{k+1} x_{k+1}}{S_{k+1}} + \sum_{i=1}^{k} \frac{\theta_i x_i}{S_{k+1}} = \frac{\theta_{k+1} x_{k+1}}{S_{k+1}} + \frac{S_k}{S_{k+1}} \sum_{i=1}^{k} \frac{\theta_i x_i}{S_k} \\
&= \frac{\theta_{k+1} x_{k+1}}{S_{k+1}} + \left(1 - \frac{\theta_{k+1}}{S_{k+1}}\right) \sum_{i=1}^{k} \frac{\theta_i x_i}{S_k} \in C,
\end{aligned}$$

since C is convex. Since $y_1 = x_1 \in C$, by induction on k, $y = y_n \in C$. ■

A2 [3] We showed in class that $conv_2(M)$ is convex. Since each point in M is a convex combination of points in M, $M \subseteq conv_2(M)$, so $conv_1(M) \subseteq conv_2(M)$. If $C \supseteq M$ is convex, then, by part A1, any convex combination of points in M is in C. Thus, $conv_2(M) \subseteq conv_1(M)$. ■

B1 [2+2] $HP(a,b)$ is convex. If $\theta \in [0,1]$ and $x_1, x_2 \in HP(a,b)$, then

$$a^T(\theta x_1 + (1-\theta)x_2) = \theta a^T x_1 + (1-\theta) a^T x_2 = \theta b + (1-\theta) b = b. \quad ■$$

If $x_1 \in HP(a, b_1)$ and $x_2 \in HP(a, b_2)$, then, by Cauchy-Schwarz,

$$\|x_1 - x_2\| \geq \left| \frac{a}{\|a\|}(x_1 - x_2) \right| = \boxed{\frac{|b_1 - b_2|}{\|a\|},}$$

and it is easily checked that $x_1 = \frac{b_1}{\|a\|^2} a$ and $x_2 = \frac{b_2}{\|a\|^2} a$ achieve this bound.

B2 [2+2] $HS(a,b)$ is convex. If $\theta \in [0,1]$ and $x_1, x_2 \in HS(a,b)$, then

$$a^T(\theta x_1 + (1-\theta)x_2) = \theta a^T x_1 + (1-\theta) a^T x_2 \leq \theta b + (1-\theta) b = b. \quad ■$$

$HS(a_1, b_1) \subseteq HS(a_2, b_2)$ if and only if $\exists c \in \mathbb{R}$ with $a_1 = c a_2$ and $b_1 \leq c b_2$.

B3 [2] $\forall x \in \mathbb{R}^d$,

$$\begin{aligned}
&\|u - x\|_2^2 \leq \|v - x\|_2^2 \\
\Leftrightarrow &\|u\|_2 - 2u^T x + \|x\|_2 \leq \|v\|_2 - 2v^T x + \|x\|_2 \\
\Leftrightarrow &\|u\| - \|v\| \leq 2(u - v)^T x.
\end{aligned}$$

Thus, $\{x \in \mathbb{R}^d \mid \|u - x\| \leq \|v - x\|\} = HS(2(u-v), \|u\| - \|v\|)$, and is thus convex. ■

C [2+3] $\forall \theta \in [0,1], x, y \in \mathbb{R}_+$,

$$f(s(\theta x + (1-\theta)y) = f(\theta sx + (1-\theta)sy) \leq \theta f(sx) + (1-\theta)f(sy). \quad \blacksquare$$

Note that, via the change of variables $u = t/x$,

$$F(x) = \frac{1}{x}\int_0^x f(t)\,dt = \frac{1}{x}\int_0^1 f(xu)x\,du = \int_0^1 f(xu)\,du.$$

Thus, $\forall \theta \in [0,1], x, y \in \mathbb{R}_+$, by convexity of the function $u \mapsto f(xu)$,

$$\begin{aligned} F(\theta x + (1-\theta)y) &= \int_0^1 f((\theta x + (1-\theta)y)u)\,du \\ &\leq \int_0^1 \theta f(xu) + (1-\theta)f(yu)\,du = \theta F(x) + (1-\theta)F(y). \quad \blacksquare \end{aligned}$$

D [3+2] The LP can be written in standard form as an LP over 6 variables:

$$0 \leq u = \begin{bmatrix} x_2 \\ y_2 \\ z_1 \\ z_2 \\ s_1 \\ s_2 \end{bmatrix}, \quad c = \begin{bmatrix} 3 \\ -1 \\ 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} \quad A = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & -1 & 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 2 \\ -1 \end{bmatrix}$$

The optimum occurs at $\boxed{(x,y,z) = (1,-1,1),}$ when $\boxed{3x - y + z = 5.}$

Shashank Singh[1]

2 LPs and gradient descent in Stats/ML [25 points] (Sashank)

A [4+4+5]

(a) Suppose β optimizes (1). Define

$$\beta_i^+ := \begin{cases} \beta_i : & \beta_i \geq 0 \\ 0 : & \text{else} \end{cases},$$

$\beta^- := \beta^+ - \beta$. Then, $y = X\beta = X(\beta^+ - \beta^-)$ and $\beta^+, \beta^- \geq 0$, so (β^+, β^-) is feasible for (2). Since $1^T(\beta^+ + \beta^-) = \sum_{i=1}^p |\beta_i| = \|\beta\|_1$, the optimum for (2) is at most $\|\beta_1\|$. ■

(b) Suppose (β^+, β^-) optimizes (2). Define $\beta := \beta^+ - \beta^-$. Then, $y = X(\beta^+ - \beta^-) = X\beta$, so β is feasible for (1). Since $\|\beta\|_1 = \sum_{i=1}^p |\beta_i| = 1^T(\beta^+ + \beta^-)$, the optimum for (1) is at most, and therefore equal to, the optimum for (1). ■

(c)

B [6+6]

(a) Rewriting in vector notation (where $h_j(x) = x_j \in \mathbb{R}^n$ is the j^{th} feature vector), we have

$$\hat{\alpha}_j = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}} \|\alpha_j h_j(x) + \hat{y} - y\|_2^2 = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}} \|\alpha_j x_j + \hat{y} - y\|_2^2 = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}} \|\alpha_j x_j - (y - \hat{y})\|_2^2,$$

from which it is apparent that $\hat{\alpha}_j$ is the length of the projection of $y - \hat{y}$ onto x_j,

$$\hat{\alpha}_j = \left\langle \frac{x_j}{\|x_j\|}, y - \hat{y} \right\rangle = \boxed{\langle x_j, y - \hat{y} \rangle.} \tag{1}$$

Note that, rewriting terms as vectors, g is a gradient of the 2-norm recentered at y:

$$g = \frac{\partial L(y, \hat{y})}{\partial \hat{y}} = \frac{\partial \|y - \hat{y}\|_2^2}{\partial \hat{y}} = 2(\hat{y} - y).$$

Thus, rewriting again in vector notation, we have

$$j = \operatorname*{argmin}_{\ell \in \{1,\dots,M\}} \| - g - \hat{\alpha}_\ell h_\ell(x)\|_2^2 = \operatorname*{argmin}_{\ell \in \{1,\dots,M\}} \|\hat{\alpha}_\ell x_\ell - (y - \hat{y})\|_2^2.$$

From (1), it is clear that this term is just the error of approximating $(y - \hat{y})$ by its projection onto x_j. This error is minimized by maximizing the inner product of x_j and $y - \hat{y}$, and hence

$$\boxed{j = \operatorname*{argmax}_{\ell \in \{1,\dots,M\}} |\langle x_j, y - \hat{y} \rangle|.} \tag{2}$$

We could make this derivation a bit more rigorous (find roots of the derivative to compute $\hat{\alpha}_j$, and then obtain (2) via some algebra), but these arguments give much better intuition.

[1] sss1@andrew.cmu.edu

(b)

$$\hat{\alpha}_j = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}} \sum_{i=1}^{n} \log\left(1 + \exp(-2y_i(\hat{y}_i + \alpha_j h_j(x_i)))\right).$$

I don't see a good way of minimizing this analytically. A simple way to approximately minimize this in practice would be to find the α_j values that minimize each term of the sum, and then try values of α_j (perhaps uniformly) in the interval surrounded by those values.

Shashank Singh[1]

3 Programming gradient descent [25 points] (Yifei)

(a) (5 pts) Based on the plots, f_Q, f_{LL}, and f_R appear convex, whereas f_H appears to have minima that are local but not global.

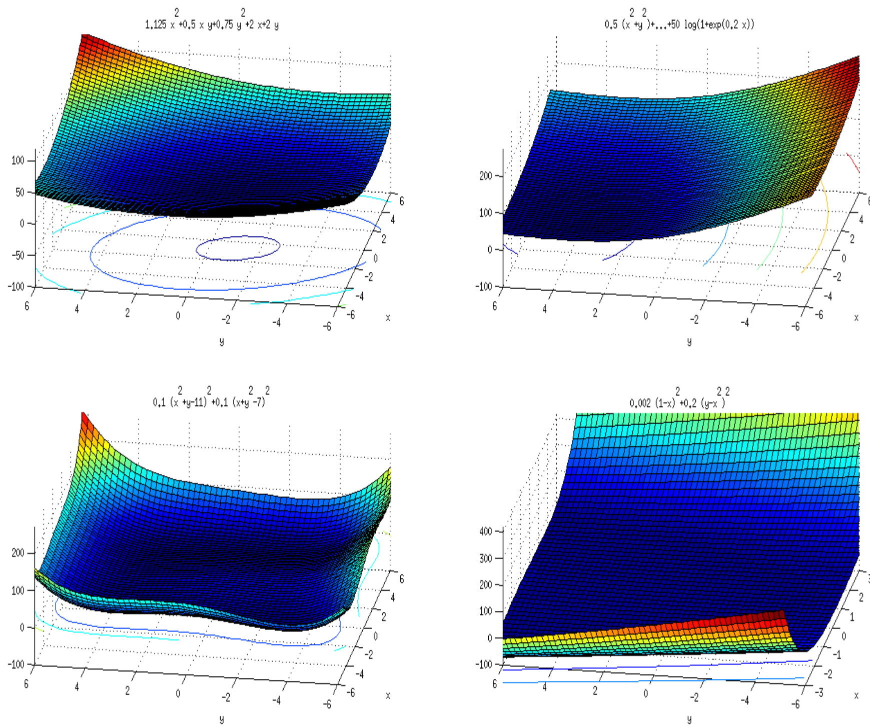

Figure 1: Surface and contour plots of the four objective functions.

(b) (8 pts) In each figure below, the row indicates the step size (0.3 or 0.8) and the column indicates the initialization ($(2,3)^T$ or random).

(c) (6 pts)

(d) (4 pts)

(e) (2 pts)

[1]sss1@andrew.cmu.edu

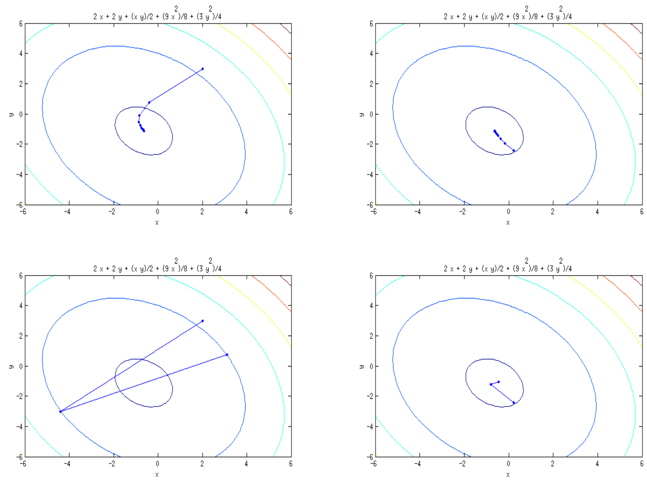

Figure 2: Gradient descent path and contour plots of f_Q at each step size and initialization.

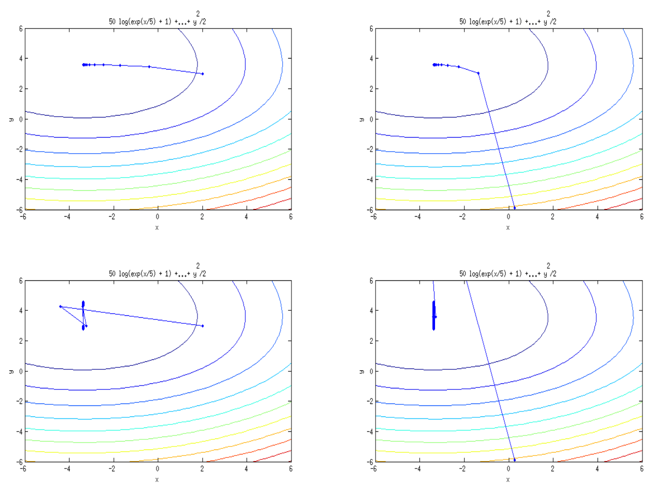

Figure 3: Gradient descent path and contour plots of f_{LL} at each step size and initialization.

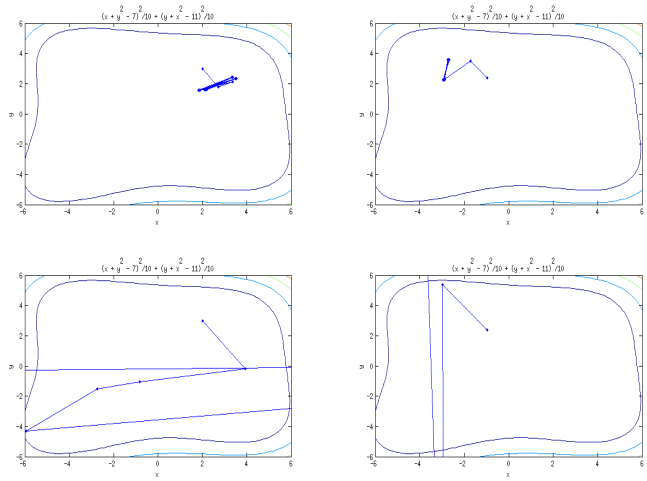

Figure 4: Gradient descent path and contour plots of f_H at each step size and initialization.

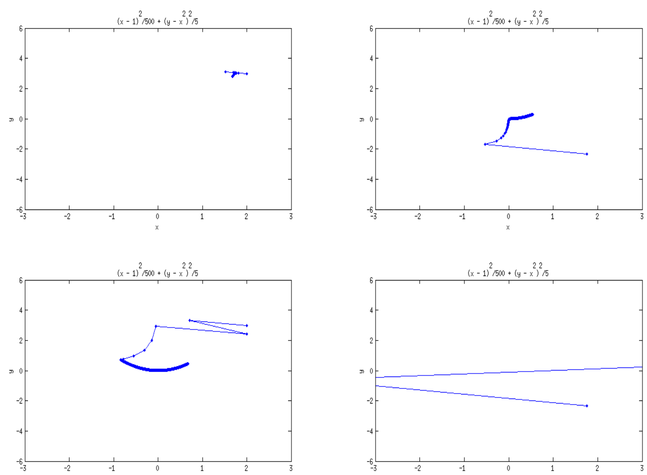

Figure 5: Gradient descent path and contour plots of f_R at each step size and initialization.

Shashank Singh[1]

4 Convergence rate of subgradient method [25 points] (Adona)

(a) (4 pts) Since the 2-norm is induced by an inner product $\langle \cdot, \cdot \rangle$,

$$\begin{aligned}
\|x^{(k)} - x^\star\|_2^2 &= \|x^{(k-1)} - x^\star - t_k g^{(k-1)}\|_2^2 && \text{(def. of } x^{(k)}\text{)} \\
&= \langle x^{(k-1)} - x^\star - t_k g^{(k-1)}, x^{(k-1)} - x^\star - t_k g^{(k-1)} \rangle \\
&= \|x^{(k-1)} - x^\star\|_2^2 - 2t_k \langle x^{(k-1)} - x^\star, g^{(k-1)} \rangle + t_k^2 \|g^{(k-1)}\|_2^2 && \text{(bilinearity of } \langle \cdot, \cdot \rangle\text{)} \\
&\leq \|x^{(k-1)} - x^\star\|_2^2 - 2t_k \left(f(x^{(k-1)}) - f(x^\star) \right) + t_k^2 \|g^{(k-1)}\|_2^2,
\end{aligned}$$

where the inequality follows from the definition of a subgradient. ■

(b) (5 pts) If g is a subgradient of f at x, then by the Lipschitz condition on f,

$$\|g\|_2^2 = g^T(x + g - x) \leq f(x+g) - f(x) \leq G\|x + g - x\|_2 = G\|g\|_2, \tag{3}$$

and so $\|g\| \leq G$. Thus, applying the recursive bound from (a) k times then gives

$$\begin{aligned}
0 \leq \|x^{(k)} - x^\star\|_2^2 &\leq \|x^{(0)} - x^\star\|_2^2 + \sum_{i=1}^k (-2t_i) \left(f(x^{(i-1)}) - f(x^\star) \right) + t_i^2 \|g^{(i-1)}\|_2^2 \\
&\leq R^2 - 2\sum_{i=1}^k t_i \left(f(x^{(i-1)}) - f(x^\star) \right) + G^2 \sum_{i=1}^k t_i^2. \quad ■
\end{aligned}$$

(c) (4 pts) Since $x_{\text{best}}^{(k)}$ is chosen so as to minimize $f(x_{\text{best}}^{(k)})$ over $\{x^{(0)}, \dots, x^{(k)}\}$,

$$2\sum_{i=1}^k t_i \left(f(x_{\text{best}}^{(k)}) - f(x^\star) \right) \leq 2\sum_{i=1}^k t_i \left(f(x^{(i-1)}) - f(x^\star) \right) \leq R^2 + G^2 \sum_{i=1}^k t_i^2,$$

using a rearrangement of the result of part (b). Thus, further rearranging, we have

$$f(x_{\text{best}}^{(k)}) - f(x^\star) \leq \frac{R^2 + G^2 \sum_{i=1}^k t_i^2}{2\sum_{i=1}^k t_i}. \quad ■ \tag{4}$$

(d) (4 pts) Plugging $t_1 = \cdots = t_k = t$ into (4) and taking the desired limit gives

$$\lim_{k\to\infty} f(x_{\text{best}}^{(k)}) - f(x^\star) \leq \lim_{k\to\infty} \frac{R^2 + G^2 k t^2}{2kt} = \boxed{\frac{G^2 t}{2}}.$$

Thus, the subgradient method with a constant step size t converges to a point at which the objective function exceeds its minimum by no more than $G^2 t/2$.

[1] sss1@andrew.cmu.edu

(e) (4 pts) Taking the desired limit in (4) gives

$$\lim_{k\to\infty} f(x_{\text{best}}^{(k)}) - f(x^\star) \le \lim_{k\to\infty} \frac{R^2 + G^2 \sum_{i=1}^k t_i^2}{2\sum_{i=1}^k t_i} \le \frac{R^2 + G^2 \lim_{k\to\infty} \sum_{i=1}^k t_i^2}{2 \lim_{k\to\infty} \sum_{i=1}^k t_i} = \boxed{0.}$$

Thus the subgradient method with step sizes as specified converges to a minimum of f.

(f) (4 pts) Plugging $t_i = R/(G\sqrt{k})$ into (4) gives

$$f(x_{\text{best}}^{(k)}) - f(x^\star) \le \frac{R^2 + R^2 k/k}{2k(R/G)\sqrt{k}} = RGk^{-3/2}. \tag{5}$$

Since the t_i was chosen to minimize (4), this is the best bound we can derive from (4).