18.335 Problem Set 5 Solutions

Problem 1: (5+5 points)

(a) Let x and y be points such that $f_i(x) \le 0$ and $f_i(y) \le 0$. Then, for any point $z = \alpha x + (1-\alpha)y$ on the line segment connecting x and y, by convexity of f_i we have $f_i(\alpha x + (1-\alpha)y) \le \alpha f_i(x) + (1-\alpha) f_i(y) \le 0$ since α and $1-\alpha$ are nonnegative. Hence every point on the line segment connecting x and y is also in the set where $f_i \le 0$, hence that set is convex.

Also, note that the intersection $C_1 \cap C_2$ of convex sets C_1, C_2 is convex, so that the intersection of all the f_i constraints gives a convex set. That is, if $x, y \in C_1 \cap C_2$, then x, y are in both C_1 and C_2, hence the line connecting x and y is in both C_1 and C_2, hence the line is in $C_1 \cap C_2$, hence $C_1 \cap C_2$ is convex.

(b) Just *finding* the feasible set becomes hard in general. For example, suppose we are solving $\min_{x\in\mathbb{R}} f_0(x)$ subject to $f_1(x) \le 0$, where the feasible set is convex—in 1d, this means it is just an interval $[a,b]$. If f_1 were a convex function, we could find feasible points from *any* starting point just by going downhill in f_1, and can in fact easily find both a and b (the edges of the feasible region) by going uphill from the minimum of f_1. However, suppose f_1 is instead an oscillatory, nonconvex function with many local minima, that just happens to be ≤ 0 only in the convex set $[a,b]$. *Finding* this convex feasible set is now hard—essentially *as hard as global optimization* of f_1, because if you start at an arbitrary infeasible point and go "downhill" you may just end up at an infeasible (> 0) local minimum.

Problem 2: (5+10+10 points)

(a) The problem is that, applying the adjoint method to compute $\frac{dg^n}{dp}$ individually for some n requires $\Theta(n)$ work to solve n steps of the adjoint recurrence for g^n. Hence doing $\Theta(n)$ work for $n = 0, \ldots, N$ is (summing the series) $\Theta(N^2)$ for G. We would like to find a $\Theta(N)$ method that *shares* work between the n's.

(b) There are potentially several ways to derive this, but one way is to look at the individual $\frac{dg^n}{dp}$ recurrences and find a way to share computations.

Imagine we applied the adjoint method for each $\frac{dg^n}{dp}$ individually as in the previous step. This involves solving an adjoint recurrence $\lambda^{n-1} = (\mathbf{f}^n_{\mathbf{x}})^T \lambda^n$ and then summing $(\lambda^i)^T \mathbf{f}^i_{\mathbf{p}}$ for $i = 1$ to n. That is, it is the *same* recurrence and the *same* final summand for each n, so why can't we just do the work once for all n's? The only difference is the initial conditions: for each n, the λ recurrence starts with $\lambda^n = (g^n_{\mathbf{x}})^T$.

The key point is to realize that the adjoint (λ) recurrence is *linear* (in λ), even if the original $\mathbf{x}$ recurrence was nonlinear. Hence, we can simply *add* the initial conditions to λ^n (at each n) to obtain a λ which is the *sum* of the solutions of the recurrences for each $\frac{dg^n}{dp}$. This insight yields the new adjoint recurrence (for G):

$$\lambda^{n-1} = (\mathbf{f}^n_{\mathbf{x}})^T \lambda^n + (g^{n-1}_{\mathbf{x}})^T,$$

$$\lambda^N = (g^N_{\mathbf{x}})^T.$$

Then we obtain

$$\frac{dG}{d\mathbf{p}} = g^0_{\mathbf{p}} + \sum_{n=1}^{N} \left[g^n_{\mathbf{p}} + (\lambda^n)^T \mathbf{f}^n_{\mathbf{p}} \right] + (\lambda^0)^T \mathbf{b}_{\mathbf{p}}.$$

(c) See attached notebook.

Problem 3: (5+5+5+10+5 points)

(a) Realize that the matrix analogue of the dot product $a^T b$ is $\operatorname{tr} A^T B = \sum_{i,j} A_{ij} B_{ij} = \operatorname{tr} BA^T$, so if we have a matrix constraint and a matrix Lagrange multiplier then we include them in trace form. Hence our Lagrangian is

$$\begin{aligned} L(E,\lambda,\Gamma) &= \operatorname{tr}(WEWE^T) + \lambda^T(E\gamma - r) + \operatorname{tr}\Gamma(E - E^T) \\ &= \boxed{\operatorname{tr}\left[WEWE^T + (E\gamma - r)\lambda^T + \Gamma(E-E^T)\right]}, \end{aligned}$$

where we have used the fact that $\lambda^T(E\gamma - r) = \operatorname{tr}\left[\lambda^T(E\gamma - r)\right] = \operatorname{tr}\left[(E\gamma - r)\lambda^T\right]$ since a scalar equals its trace.

(b) Discarding $\Theta(\Delta^2)$ terms, and using $\operatorname{tr} A = \operatorname{tr} A^T$ with $\operatorname{tr} AB = \operatorname{tr} BA$ along with the fact that W was assumed symmetric, we obtain:

$$\begin{aligned} L(E+\Delta) - L(E) &\approx \operatorname{tr}\left[W\Delta WE^T + WEW\Delta^T + \Delta\gamma\lambda^T + \Gamma(\Delta - \Delta^T)\right] \\ &= \operatorname{tr}\Delta^T\left[2WEW + \lambda\gamma^T + \Gamma^T - \Gamma\right] = 0 \qquad \text{for all } \Delta. \end{aligned}$$

Now, we have an equation of the form $\operatorname{tr}\Delta^T X = 0$ for all $\Delta \in \mathbb{R}^{N\times N}$, which immediately implies that $X = 0$, since otherwise we could choose $\Delta = X$ and get a nonzero result. Equivalently, by comparison with the Taylor series we see that X here is the "gradient" $\partial L/\partial E$, and hence we set

$$\frac{\partial L}{\partial E} = 2WEW + \lambda\gamma^T + \Gamma^T - \Gamma = 0$$

to find:

$$\boxed{E = -\frac{1}{2}W^{-1}\left(\lambda\gamma^T + \Gamma^T - \Gamma\right)W^{-1}}.$$

(c) Applying the constraint $E = E^T$, we find $\lambda\gamma^T + \Gamma^T - \Gamma = \gamma\lambda^T + \Gamma - \Gamma^T$, or $\Gamma^T - \Gamma = \frac{\gamma\lambda^T - \lambda\gamma^T}{2}$. Plugging this in above, we get:

$$\boxed{E = -\frac{1}{4}W^{-1}\left(\gamma\lambda^T + \lambda\gamma^T\right)W^{-1}},$$

which is manifestly symmetrical.

(d) Plugging this E into $E\gamma = r$ and multiplying both sides by $-4W$, we get

$$\gamma\lambda^T W^{-1}\gamma + \lambda\gamma^T W^{-1}\gamma = -4Wr \implies \lambda = -\frac{4Wr + \gamma\lambda^T W^{-1}\gamma}{\gamma^T W^{-1}\gamma}$$

and hence if we take the transpose to get $\lambda^T = \cdots$ and multiply both sides by $W^{-1}\gamma$, we get:

$$\lambda^T W^{-1}\gamma = -\frac{4r^T\gamma + \left(\lambda^T W^{-1}\gamma\right)\gamma^T W^{-1}\gamma}{\gamma^T W^{-1}\gamma},$$

which we can solve for the scalar quantity $\lambda^T W^{-1}\gamma$:

$$\lambda^T W^{-1}\gamma = -2\frac{r^T\gamma}{\gamma^T W^{-1}\gamma}.$$

Then we can plug this into $\lambda = \cdots$ to solve for λ:

$$\lambda = -\frac{4Wr - 2\gamma\frac{r^T\gamma}{\gamma^T W^{-1}\gamma}}{\gamma^T W^{-1}\gamma} = \frac{2\gamma\gamma^T r}{(\gamma^T W^{-1}\gamma)^2} - \frac{4Wr}{\gamma^T W^{-1}\gamma}.$$

Finally, we can plug this λ into $E = \cdots$ to find E:

$$
\begin{aligned}
E &= -\frac{1}{2} W^{-1} \left[\frac{\gamma\gamma^T r\gamma^T + \gamma r^T \gamma\gamma^T}{(\gamma^T W^{-1}\gamma)^2} - 2\frac{W r\gamma^T + \gamma r^T W}{\gamma^T W^{-1}\gamma} \right] W^{-1} \\
&= \frac{1}{\gamma^T W^{-1}\gamma} \left[r\gamma^T W^{-1} + W^{-1}\gamma r^T - \frac{\gamma^T r}{\gamma^T W^{-1}\gamma} W^{-1}\gamma\gamma^T W^{-1} \right].
\end{aligned}
$$

(e) If we choose $W = H^{(n+1)}$ and apply the secant condition $W^{-1}\gamma = \delta$, we get

$$
E = \frac{1}{\gamma^T \delta} \left[r\delta^T + \delta r^T - \frac{\gamma^T r}{\gamma^T \delta} \delta\delta^T \right]
$$

which is a BFGS update for $\left[H^{(n)}\right]^{-1}$: if we plug in $r = \delta - \left[H^{(n)}\right]^{-1}\gamma$, we immediately (after trivial algebra) get the formula for $E = \left[H^{(n+1)}\right]^{-1} - \left[H^{(n)}\right]^{-1}$ that was given in the problem.