Solutions

Problem 1 (Convex functions and convex sets)

(a) **i) Polyhedron.** Let $P = \{x \in \mathbb{R}^n \mid Ax \leq b\}$ (componentwise). Take any $x_1, x_2 \in P$ and any $\theta \in [0,1]$. Then

$$A(\theta x_1 + (1-\theta)x_2) = \theta A x_1 + (1-\theta)Ax_2 \leq \theta b + (1-\theta)b = b.$$

Hence $\theta x_1 + (1-\theta)x_2 \in P$, so P is convex.

ii) ℓ_1-ball. Let $B_1(r) = \{x \in \mathbb{R}^n : \|x\|_1 \leq r\}$. For any $x_1, x_2 \in B_1(r)$ and $\theta \in [0,1]$, by triangle inequality and homogeneity of $\|\cdot\|_1$,

$$\|\theta x_1 + (1-\theta)x_2\|_1 \leq \|\theta x_1\|_1 + \|(1-\theta)x_2\|_1 = \theta\|x_1\|_1 + (1-\theta)\|x_2\|_1 \leq r.$$

So $B_1(r)$ is convex.

iii) PSD cone. Let $S^n_+ = \{X \in S^n : X \succeq 0\}$. Take $X, Y \succeq 0$ and $\theta \in [0,1]$. For any $v \in \mathbb{R}^n$,

$$v^\top(\theta X + (1-\theta)Y)v = \theta v^\top X v + (1-\theta)v^\top Y v \geq 0.$$

Thus $\theta X + (1-\theta)Y \succeq 0$ and S^n_+ is convex.

(b) **i) $f(x) = \|x\|_1$ is convex.** For any $x_1, x_2 \in \mathbb{R}^n$ and $\theta \in [0,1]$,

$$\|\theta x_1 + (1-\theta)x_2\|_1 \leq \theta\|x_1\|_1 + (1-\theta)\|x_2\|_1,$$

by the same argument as in (a-ii). Hence f is convex.

ii) If f, g are convex, then $f+g$ is convex. For any x_1, x_2 and $\theta \in [0,1]$,

$$\begin{aligned}(f+g)(\theta x_1 + (1-\theta)x_2) &= f(\theta x_1 + (1-\theta)x_2) + g(\theta x_1 + (1-\theta)x_2) \\ &\leq \theta f(x_1) + (1-\theta)f(x_2) + \theta g(x_1) + (1-\theta)g(x_2) \\ &= \theta(f+g)(x_1) + (1-\theta)(f+g)(x_2).\end{aligned}$$

iii) A counterexample for composition. Let outer $f(u) = -u$ (affine, hence convex) and inner $g(x) = x^2$ (convex). Then $(f \circ g)(x) = -x^2$, which is *not* convex. So composition of convex functions is not always convex.

Problem 2 (Strong Convexity)

A differentiable function is μ-strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2, \quad \forall x, y.$$

(a) Let $x^\star$ be a minimizer of f. Since f is differentiable and convex, $\nabla f(x^\star) = 0$. Apply strong convexity with $x = x^\star$, $y = x$:

$$f(x) \geq f(x^\star) + \langle \nabla f(x^\star), x - x^\star \rangle + \frac{\mu}{2}\|x - x^\star\|_2^2 = f(x^\star) + \frac{\mu}{2}\|x - x^\star\|_2^2,$$

hence $f(x) - f(x^\star) \geq \frac{\mu}{2}\|x - x^\star\|_2^2$.

(b) For $f(x) = \frac{1}{2}x^\top Ax + b^\top x$, the Hessian is A. The largest valid μ equals $\lambda_{\min}(A)$. Given

$$A = \begin{pmatrix} 6 & 2 \\ 2 & 9 \end{pmatrix},$$

its eigenvalues satisfy $\lambda^2 - \mathrm{tr}(A)\lambda + \det(A) = 0$:

$$\lambda^2 - 15\lambda + 50 = 0 \;\Rightarrow\; \lambda = \frac{15 \pm \sqrt{225 - 200}}{2} = \frac{15 \pm 5}{2} \in \{10, 5\}.$$

Thus $\mu_{\max} = \lambda_{\min}(A) = 5$.

Problem 3 (Prox operators for ℓ_0 and ℓ_2)

(a) $\mathrm{prox}_{\lambda\|\cdot\|_0}(v)$. Recall $\|x\|_0 = \sum_{i=1}^n \mathbf{1}(x_i \neq 0)$. The objective is separable:

$$\begin{aligned}\mathrm{prox}_{\lambda\|\cdot\|_0}(v) &= \arg\min_{x\in\mathbb{R}^n}\left\{\lambda\|x\|_0 + \tfrac{1}{2}\|x - v\|_2^2\right\} \\ &= \left(\arg\min_{x_i\in\mathbb{R}}\{\lambda\mathbf{1}(x_i \neq 0) + \tfrac{1}{2}(x_i - v_i)^2\}\right)_{i=1}^n.\end{aligned}$$

Fix coordinate i. Consider two cases:

If $x_i = 0$: $\quad \lambda\mathbf{1}(x_i \neq 0) + \frac{1}{2}(x_i - v_i)^2 = \frac{1}{2}v_i^2$.

If $x_i \neq 0$: $\quad \lambda + \frac{1}{2}(x_i - v_i)^2$ is minimized at $x_i = v_i$, giving value λ.

Hence we choose $x_i = v_i$ if $\lambda < \frac{1}{2}v_i^2$, i.e., $|v_i| > \sqrt{2\lambda}$, and choose $x_i = 0$ if $\lambda > \frac{1}{2}v_i^2$, i.e., $|v_i| < \sqrt{2\lambda}$. At the tie $|v_i| = \sqrt{2\lambda}$, both 0 and v_i are minimizers. Therefore,

$$\left(\mathrm{prox}_{\lambda\|\cdot\|_0}(v)\right)_i \in \begin{cases}\{v_i\}, & |v_i| > \sqrt{2\lambda}, \\ \{0, v_i\}, & |v_i| = \sqrt{2\lambda}, \\ \{0\}, & |v_i| < \sqrt{2\lambda}.\end{cases}$$

Equivalently (choosing the common convention that breaks ties by 0), this is the hard-thresholding rule

$$\left(\mathrm{prox}_{\lambda\|\cdot\|_0}(v)\right)_i = \begin{cases} v_i, & |v_i| > \sqrt{2\lambda}, \\ 0, & |v_i| \leq \sqrt{2\lambda}.\end{cases}$$

(b) $\text{prox}_{\lambda\|\cdot\|_2}(v)$. We solve

$$\text{prox}_{\lambda\|\cdot\|_2}(v) = \arg\min_{x\in\mathbb{R}^n}\Big\{\lambda\|x\|_2 + \tfrac{1}{2}\|x-v\|_2^2\Big\}.$$

Write $x = ru$ where $r = \|x\|_2 \geq 0$ and $\|u\|_2 = 1$ (for $x \neq 0$). Then

$$\|x-v\|_2^2 = \|ru-v\|_2^2 = r^2 - 2r\langle u, v\rangle + \|v\|_2^2.$$

For fixed r, minimizing over $\|u\|_2 = 1$ is equivalent to maximizing $\langle u, v\rangle$, achieved by $u = \frac{v}{\|v\|_2}$ when $v \neq 0$, giving $\max\langle u, v\rangle = \|v\|_2$. Thus the problem reduces to

$$\min_{r\geq 0}\ \lambda r + \tfrac{1}{2}(r - \|v\|_2)^2 \quad \text{(up to an additive constant } \tfrac{1}{2}\|v\|_2^2\text{)}.$$

Solve this quadratic problem coordinate by coordinate we can get

$$r^\star = \max(\|v\|_2 - \lambda,\, 0).$$

Hence, for $v \neq 0$,

$$x^\star = r^\star\frac{v}{\|v\|_2} = \begin{cases}\left(1 - \dfrac{\lambda}{\|v\|_2}\right)v, & \|v\|_2 > \lambda,\\ 0, & \|v\|_2 \leq \lambda.\end{cases}$$

This is the vector (group) soft-thresholding / shrinkage operator:

$$\text{prox}_{\lambda\|\cdot\|_2}(v) = \left(1 - \frac{\lambda}{\|v\|_2}\right)_+ v, \quad \text{where } (t)_+ = \max(t, 0),$$

and the formula also yields 0 when $v = 0$.

(c) Compute three proximal operators for $\lambda = 1$.
Let $v^{(1)} = (1, 1)$ and $v^{(2)} = (2, 0)$, with $\lambda = 1$ so $\sqrt{2\lambda} = \sqrt{2}$.

$$\begin{aligned}&\text{prox}_{\|\cdot\|_0}\big(v^{(1)}\big) = (0,0) \quad \text{since } |1| \leq \sqrt{2} \text{ for both coordinates},\\ &\text{prox}_{\|\cdot\|_0}\big(v^{(2)}\big) = (2,0) \quad \text{since } |2| > \sqrt{2},\ |0| \leq \sqrt{2}.\end{aligned}$$

$$\begin{aligned}\text{prox}_{\|\cdot\|_1}\big(v^{(1)}\big) &= \big(\max(1-1,0), \max(1-1,0)\big) = (0,0),\\ \text{prox}_{\|\cdot\|_1}\big(v^{(2)}\big) &= \big(\max(2-1,0), 0\big) = (1,0).\end{aligned}$$

$$\begin{aligned}\text{prox}_{\|\cdot\|_2}\big(v^{(1)}\big) &= \left(1 - \frac{1}{\|(1,1)\|_2}\right)(1,1) = \left(1 - \frac{1}{\sqrt{2}}\right)(1,1),\\ \text{prox}_{\|\cdot\|_2}\big(v^{(2)}\big) &= \left(1 - \frac{1}{\|(2,0)\|_2}\right)(2,0) = \left(1 - \frac{1}{2}\right)(2,0) = (1,0).\end{aligned}$$

Problem 4 (Convergence of Projected Gradient Descent)

(a) Since f is L-smooth, for all x, y we have the descent lemma

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

Taking $x = x_k$ and $y = x_{k+1}$ gives

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

Adding $g(x_{k+1})$ to both sides and using the definition of $Q_L(\cdot; x_k)$ yields

$$F(x_{k+1}) = f(x_{k+1}) + g(x_{k+1}) \le Q_L(x_{k+1}; x_k).$$

(b) The function $x \mapsto \frac{L}{2}\|x - x_k\|^2$ is L-strongly convex, the term $x \mapsto f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$ is affine (hence convex), and g is convex. Therefore their sum $Q_L(\cdot; x_k)$ is L-strongly convex.

Let $x_{k+1} = \arg\min_x Q_L(x; x_k)$. A standard consequence of μ-strong convexity (with $\mu = L$) is: for all x,

$$Q_L(x; x_k) \ge Q_L(x_{k+1}; x_k) + \frac{L}{2}\|x - x_{k+1}\|^2.$$

This is exactly the desired inequality.

(c) From (a) and (b), for any $x \in \mathbb{R}^n$,

$$F(x_{k+1}) \le Q_L(x_{k+1}; x_k) \le Q_L(x; x_k) - \frac{L}{2}\|x - x_{k+1}\|^2,$$

hence

$$F(x_{k+1}) \le Q_L(x; x_k) - \frac{L}{2}\|x - x_{k+1}\|^2, \qquad \forall x.$$

Next, by convexity of f,

$$f(x) \ge f(x_k) + \langle \nabla f(x_k), x - x_k \rangle,$$

so

$$\begin{aligned} Q_L(x; x_k) &= f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 + g(x) \\ &\le f(x) + g(x) + \frac{L}{2}\|x - x_k\|^2 \\ &= F(x) + \frac{L}{2}\|x - x_k\|^2. \end{aligned}$$

Substituting this bound into the previous inequality yields the key estimate:

$$F(x_{k+1}) \le F(x) + \frac{L}{2}\Big(\|x - x_k\|^2 - \|x - x_{k+1}\|^2\Big), \qquad \forall x.$$

(d) Let $x^\star \in X^\star$ be any minimizer, and denote $F^\star = F(x^\star) = \min_x F(x)$. Applying the key estimate in (c) with $x = x^\star$ gives, for all k,

$$F(x_{k+1}) - F^\star \le \frac{L}{2}\Big(\|x^\star - x_k\|^2 - \|x^\star - x_{k+1}\|^2\Big).$$

Summing from $k = 0$ to $K-1$ (telescoping) yields

$$\sum_{k=0}^{K-1} \big(F(x_{k+1}) - F^\star\big) \le \frac{L}{2}\Big(\|x^\star - x_0\|^2 - \|x^\star - x_K\|^2\Big) \le \frac{L}{2}\|x_0 - x^\star\|^2.$$

Moreover, choosing $x = x_k$ in (c) gives monotonic decrease:

$$F(x_{k+1}) \le F(x_k), \quad \forall k,$$

so $F(x_K) \le F(x_k)$ for all $k \le K$, and hence

$$F(x_K) - F^\star \le \frac{1}{K}\sum_{k=0}^{K-1} \big(F(x_{k+1}) - F^\star\big) \le \frac{L}{2K}\|x_0 - x^\star\|^2.$$

This proves the $O(1/K)$ convergence rate of objective values.

Problem 5 (Backtracking when L is unknown)

Backtracking rule: accept $\widehat{L}$ if

$$f\left(x - \frac{1}{\widehat{L}}\nabla f(x)\right) \le f(x) - \frac{1}{2\widehat{L}}\|\nabla f(x)\|_2^2.$$

(a) Let $y = x - \frac{1}{\widehat{L}}\nabla f(x)$. By descent lemma (with true L),

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2.$$

Here $y - x = -\frac{1}{\widehat{L}}\nabla f(x)$, so

$$f(y) \le f(x) - \frac{1}{\widehat{L}}\|\nabla f(x)\|_2^2 + \frac{L}{2}\cdot\frac{1}{\widehat{L}^2}\|\nabla f(x)\|_2^2 = f(x) - \left(\frac{1}{\widehat{L}} - \frac{L}{2\widehat{L}^2}\right)\|\nabla f(x)\|_2^2.$$

If $\widehat{L} \ge L$, then

$$\frac{1}{\widehat{L}} - \frac{L}{2\widehat{L}^2} = \frac{1}{2\widehat{L}}\left(2 - \frac{L}{\widehat{L}}\right) \ge \frac{1}{2\widehat{L}},$$

hence

$$f\left(x - \frac{1}{\widehat{L}}\nabla f(x)\right) \le f(x) - \frac{1}{2\widehat{L}}\|\nabla f(x)\|_2^2.$$

Therefore any $\widehat{L} \ge L$ will be accepted, so the loop terminates in finitely many trials.

(b) Starting from $\widehat{L} = \widehat{L}_k$, after N multiplications the trial value is $\widehat{L}_k\eta^N$. The loop must accept once $\widehat{L}_k\eta^N \ge L$. Thus the smallest such N satisfies

$$N \le \left\lceil \log_\eta\left(\frac{L}{\widehat{L}_k}\right)\right\rceil_+, \quad \text{where } \lceil t\rceil_+ = \max\{0, \lceil t\rceil\}.$$

Problem 6 (APGD objective is not necessarily monotone)

APGD update:

$$y_k = x_k + \beta(x_k - x_{k-1}), \qquad x_{k+1} = y_k - \alpha \nabla f(y_k).$$

Take $f(x) = \frac{1}{2}x^2$, $\nabla f(x) = x$, $\alpha = \frac{1}{2}$, $\beta = 4$, and $x_{-1} = x_0 = 1$.

(a) First step:

$$y_0 = x_0 + 4(x_0 - x_{-1}) = 1, \quad x_1 = y_0 - \tfrac{1}{2}y_0 = \tfrac{1}{2}.$$

Second step:

$$y_1 = x_1 + 4(x_1 - x_0) = \tfrac{1}{2} + 4\left(\tfrac{1}{2} - 1\right) = \tfrac{1}{2} - 2 = -\tfrac{3}{2},$$
$$x_2 = y_1 - \tfrac{1}{2}y_1 = \tfrac{1}{2}y_1 = -\tfrac{3}{4}.$$

Compute objective values:

$$f(x_1) = \tfrac{1}{2}\left(\tfrac{1}{2}\right)^2 = \tfrac{1}{8}, \qquad f(x_2) = \tfrac{1}{2}\left(\tfrac{3}{4}\right)^2 = \tfrac{9}{32}.$$

Since $\frac{9}{32} > \frac{1}{8}$, the objective increases, so APGD is not necessarily monotone.

(b) The extrapolation $y_k = x_k + \beta(x_k - x_{k-1})$ adds "momentum" in the recent moving direction. If β is large, y_k can jump past the minimizer (overshoot), so the next gradient step may land at a point with a larger function value.