18.657 PS 3 SOLUTIONS

1. Problem 1

(1) Symmetry is clear. Let K_1 and K_2 be the PSD Gram matrices of k_1 and k_2, respectively. Then the Gram matrix K of k is simply the Hadamard (or Schur) product $K_1 \bullet K_2$; we wish to see that this is PSD.

The Kronecker product $K_1 \otimes K_2$ is PSD: its eigenvalues are simply the pair products of an eigenvalue of K_1 with an eigenvalue from K_2, as is easily seen from the identity

$$(K_1 \otimes K_2)(v \otimes w) = (K_1 v) \otimes (K_2 w)$$

when v and w are eigenvectors of K_1 and K_2, respectively. Now the Hadamard product $K = K_1 \bullet K_2$ is a principal submatrix of $A \otimes B$, and a principal submatrix of a PSD matrix is PSD.

(There are many good approaches to this part.)

(2) Treating g as a real vector indexed by $\mathcal{C}$, we have $K = gg^\top$, and a matrix of this form is always PSD.

(3) The Gram matrix K of k is simply $Q(K_1)$, where K_1 is the Gram matrix of k_1, and where the multiplication in the polynomial is Hadamard product. From (1), the PSD matrices are closed under Hadamard product, and it is a common fact that they are closed under positive scaling and addition; thus they are closed under the application of polynomials with non-negative coefficients.

(4) Let $T_r(x)$ be the rth Taylor approximation to $\exp(x)$ about 0, a polynomial with non-negative coefficients. Then $T_r(k_1)$ converges to $k = \exp(k_1)$ as $r \to \infty$; equivalently, $T_r(K_1)$ converges to $K = \exp(K_1)$, where K and K_1 are the Gram matrices of k and k_1. Here exp is the entry-wise exponential function, not the "matrix exponential".. From (3), $T_r(K_1)$ is PSD. As the PSD cone is a closed subset of $\mathbb{R}^{n\times n}$ (it's defined by non-strict inequalities $x^\top M x \geq 0$), the limit K is PSD, and this is the Gram matrix of K.

(5) Applying (2) to the function $f(u) = \exp(-\|u\|^2)$, the kernel $k_1(u,v) = \exp(-\|u\|^2 - \|v\|^2)$ is PSD. Moreover the kernel $k_2(u,v) = 2u^\top v$ is PSD: if $\mathcal{C}$ is a set of vectors, and we let the matrix U be defined by taking the elements of $\mathcal{C}$ as columns, then the Gram matrix of k_2 is $2U^\top U$ which is PSD. By (4), the kernel $k_3(u,v) = \exp(2u^\top v)$ is PSD. By (1), the kernel

$$k(u,v) = k_1(u,v)k_3(u,v) = \exp(-\|u\|^2 + 2u^\top v - \|v\|^2)\exp(-\|u-v\|^2)$$

is PSD.

Date: Fall 2015.

2. Problem 2

(1) Suppose $x \in \mathbb{R}^d$; we wish to find $y \in C$ minimizing $\|x - y\|^2 = \sum_{i=1}^n (x_i - y_i)^2$. As the constraints defining C apply to each y_i separately, the problem amounts to finding, for each i, a value $-1 \le y_i \le 1$ minimizing $(x_i - y_i)^2$. This is clearly achieved at

$$y_i = \begin{cases} x_i/|x_i| & \text{if } |x_i| > 1, \\ x_i & \text{otherwise.} \end{cases}$$

This formula is exact, so there is no convergence issue; effectively the method converges perfectly after one update on each coordinate.

(2) Let $z \in \mathbb{R}^d$ be given; we want $x \in \Delta$ minimizing the ℓ_2 distance $f(x) = \|x - z\|$. We apply mirror descent, following the corollary on page 7 of the Lecture 13 notes. The objective is clearly 1-Lipschitz, with gradient $\nabla f(x) = (x - z)/\|x - z\|$, so we obtain the following convergence guarantee at iteration k:

$$f(x_k^\circ) - f(x^*) \le \sqrt{\frac{2 \log d}{k}}.$$

(3) (a) $\S_n^+$ is convex: certainly any convex combination of symmetric matrices is symmetric, and if $A, B \in \S_n^+$, then

$$x^\top(\lambda A + (1-\lambda)B)x = \lambda x^\top A x + (1-\lambda) x^\top B x$$

is a convex combination of non-negative reals, thus non-negative, so a convex combination of PSD matrices is PSD.

$\S_n^+$ is closed in $\mathbb{R}^{n \times n}$ as it is defined as an intersection of a linear subspace (the symmetric matrices) and the half-spaces $\langle M, v^\top v\rangle \ge 0$ for $v \in \mathbb{R}^n$, all of which are closed; an intersection of closed sets is closed.

(b) Let $A \in \S_n$. As $\S_n^+$ is convex and closed, and the function $f(B) = \frac{1}{2}\|A - B\|_F^2$ is convex, a matrix B minimizes f over S_n^+ iff it satisfies first-order optimality. Specifically, we must have that $\nabla f(B) = \sum_i \mu_i x_i x_i^\top$, for some $\mu_i \ge 0$ and some x_i for which B is tight for the constraint $x_i^\top B x_i \ge 0$ defining S_n^+.

We compute the gradient:

$$\nabla(\frac{1}{2}\|A - B\|_F^2) = \nabla \frac{1}{2} \operatorname{tr}(A^2 - 2AB + B^2) = B - A.$$

Thus, if we can write $B - A = \sum_i \mu_i x_i x_i^\top$ as above, then we certify B as optimal.

Write the eigendecomposition $A = U \Sigma U^\top$, and let $B = U \Sigma_+ U^\top$, where Σ_+ replaces the negative entries of Σ by zero. Then

$$B - A = U(\Sigma_+ - \Sigma)U^\top = \sum_{i : \Sigma_{ii} < 0} (-\Sigma_{ii}) U_i U_i^\top,$$

and we have $U_i^\top B U_i = (\Sigma_+)_{ii} = 0$, thus fulfilling first-order optimality.

(4) We begin with the first claim. In class, we proved that

$$\langle \pi(x) - x, \pi(x) - z\rangle \le 0,$$

whenever $z \in \mathcal{C}$. Applying this with $z = \pi(y)$, we have

$$\langle \pi(x) - x, \pi(x) - \pi(y)\rangle \le 0.$$

Summing a copy of this inequality with the same thing with x and y reversed, we have

$$\langle \pi(x) - \pi(y) - (x - y), \pi(x) - \pi(y)\rangle \le 0,$$

$$\pi(x) - \pi(y)\|^2 \le \langle x - y, \pi(x) - \pi(y)\rangle \le \|x - y\|\|\pi(x) - \pi(y)\|,$$

by Cauchy–Schwartz. Cancelling $\|\pi(x) - \pi(y)\|$ from both sides yields the first claim.

The second claim follows by specializing to the case $y \in \mathcal{C}$, for which $\pi(y) = y$.

3. Problem 3

(1) (a) By definition, $f^*(y) = \sup_{x>0} xy - \frac{1}{x}$. When $y > 0$, a sufficiently large choice of x makes the objective arbitrarily large, and the supremum is infinite. When $y \le 0$, the objective is bounded above by 0; for $y = 0$, this is achieved as $x \to \infty$, whereas for $y < 0$, first-order optimality conditions show that the optimum is achieved at $x = (-y)^{-1/2}$, at which $f^*(y) = -2\sqrt{-y}$. We have $D = (-\infty, 0]$.

(b) By definition, $f^*(y) = \sup_{x\in\mathbb{R}^d} y^\top x - \frac{1}{2}|x|_2^2$. The gradient of the objective is $y - x$, so first-order optimality conditions are satisfied at $x = y$, and we have $f^*(y) = \frac{1}{2}|x|_2^2$ (f is self-conjugate). Here $D = \mathbb{R}^d$.

(c) By definition, $f^*(y) = \sup_{x\in\mathbb{R}^d} y^\top x - \log \sum_{j=1}^d \exp(x_j)$. The partial derivative in x_i of the objective function is

$$y_i - \frac{\exp x_i}{\sum_j \exp x_j},$$

so the gradient may be made zero whenever y lies in the simplex Δ, by taking $x_i = \log y_i$. For such y, we thus have $f^*(y) = \sum_i y_i \log y_i$.

We next rule out all $y \notin \Delta$, so that $D = \Delta$. Consider x of the form $(\lambda, \lambda, \dots, \lambda)$, for which the objective value is $\lambda \sum_i y_i - \lambda - \log d$. When $\sum_i y_i \ne 1$, this can be made arbitrarily large, by taking an extreme value of λ. On the other hand, if any coordinate y_i of y is negative, then taking x to be supported only on the ith coordinate, the objective value is $y_i x_i - \log((d-1) + \exp x_i)$, which is arbitrarily large when we take x_i to be a sufficiently large negative number. So y must lie in the simplex Δ.

(2) For all $x \in C$ and $y \in D$, we have $f^*(y) \ge y^\top x - f(x)$, so that $y^\top x - f^*(y) \le f(x)$. Thus $f^{**}(x) = \sup_{y\in D} y^\top x - f^*(y) \le f(x)$.

(3) Here $C = \mathbb{R}^d$, so that the supremum is either achieved in some limit of arbitrarily distant points, or else at at point satisfying first-order optimality. The first case can actually occur, e.g. when $d = 1$, $f(x) = -\exp(x)$, and $y = 0$. In the other case, first-order optimality is that

$$0 = \nabla_x(y^\top x - f(x)) = y - \nabla f(x),$$

so that $\nabla f(x^*) = y$.

(4) We will need the gradient of f^*. As f is strictly convex, $y = \nabla f(x)$ is an injective function of x, so we can write $x = (\nabla f)^{-1}(y)$. Then

$$\begin{aligned}
f^*(y) &= y^\top x^* - f(x^*) \\
&= y^\top (\nabla f)^{-1}(y) - f((\nabla f)^{-1}(y)), \\
\nabla f^*(y) &= (\nabla f)^{-1}(y) + D(\nabla f)^{-1}(y)[y] - D(\nabla f)^{-1}(y)[\nabla f((\nabla f)^{-1})(y))] \\
&= (\nabla f)^{-1}(y),
\end{aligned}$$

where $Df(a)[b]$ denotes the Jacobian of f, taken at a, applied to the vector b.

We now compute the Bregman divergence:

$$\begin{aligned}
D_{f^*}(\nabla f(y), \nabla f(x)) &= f^*(\nabla f(y)) - f^*(\nabla f(x)) - (\nabla f^*(\nabla f(x)))^\top (\nabla f(y) - \nabla f(x)) \\
&= \nabla f(y)^\top y - f(y) - \nabla f(x)^\top x + f(x) - x^\top (\nabla f(y) - \nabla f(x)) \\
&= f(x) - f(y) - (\nabla f(y))^\top (x - y) \\
&= D_f(x, y).
\end{aligned}$$

4. Problem 4

(1) Adapting the proof of convergence of projected subgradient descent from the lecture notes, we have:

$$\begin{aligned}f(x_s) - f(x^*) &\le g_s^\top (x_s - x^*) \\ &= \frac{\|g_s\|}{\eta}(x_s - y_{s+1})^\top (x_s - x^*) \\ &\le \frac{\|g_s\|}{2\eta}(\|x_s - y_{s+1}\|^2 + \|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2) \\ &= \frac{\eta\|g_s\|}{2} + \frac{\|g_s\|}{2\eta}(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2),\end{aligned}$$

as $x_s - y_{s+1}$ has norm η,

$$\le \frac{\eta\|g_s\|}{2} + \frac{\|g_s\|}{2\eta}(\|x_s - s^*\|^2 - \|x_{s+1} - x^*\|^2),$$

as the projection operator is a contraction.

Summing over s, and using the bounds $\|x_1 - x^*\| \le R$ and $\|g_s\| \le L$, we have

$$f(\bar{x}_s) - f(x^*) \le \frac{\eta L}{2} + \frac{LR^2}{2\eta k}.$$

Taking $\eta = R/\sqrt{k}$, we obtain the rate $LR/\sqrt{k}$.

(2) (a) Starting from the definition of β-smoothness:

$$\begin{aligned}f(x_{s+1}) - f(x_s) &\le \nabla f(x_s)^\top (x_{s+1} - x_s) + \frac{\beta}{2}\|x_{s+1} - x_s\|^2 \\ &= \gamma_s \nabla f(x_s)^\top (y_s - x_s) + \frac{\beta}{2}\gamma_s^2\|y_s - x_s\|^2 \\ &\le \gamma_s \nabla f(x_s)^\top (x^* - x_s) + \frac{\beta}{2}\gamma_s^2 R^2,\end{aligned}$$

as $f(x_s)^\top y_s \le f(x_s)^\top y$ for all $y \in C$,

$$\le \gamma_s(f(x^*) - f(x_s)) + \frac{\beta}{2}\gamma_s^2 R^2,$$

by convexity.

(b) We induct on k. Continuing from the inequality above by subtracting $f(x^*) - f(x_s)$ from both sides, we have

$$f(x_{s+1}) - f(x^*) \le (1 - \gamma_s)(f(x_s) - f(x^*)) + \frac{\beta}{2}\gamma_s^2 R^2,$$

or, if we define $\delta_s = f(x_s) - f(x^*)$,

$$\delta_{s+1} \le (1 - \gamma_s)\delta_s + \frac{\beta}{2}\gamma_s^2 R^2.$$

Specializing to $s = 1$, and noting that $\gamma_1 = 1$, we have that $\delta_2 = \beta R^2/2 \le 2\beta R^2/3$, so that the base case of $k = 2$ is satisfied. Now proceeding inductively, we have

$$\begin{aligned}\delta_k &\le (1-\gamma_{k-1})\delta_{k-1} + \frac{\beta}{2}\gamma_{k-1}^2 R^2 \\ &\le \frac{2(k-2)\beta R^2}{k^2} + \frac{2\beta R^2}{k^2},\end{aligned}$$

by induction and the definition of γ_{k-1},

$$\begin{aligned}&= \frac{2(k-1)\beta R^2}{k^2} \\ &\le \frac{2\beta R^2}{k+1},\end{aligned}$$

completing the induction.