10702 Homework 2 Solution

Thanks to Akshay Krishnamurthy for providing his solution.

1 Convexity and Optimization

1. (Convexity)

   (a) We'll show that the second derivative of $1/g(x)$ is always positive, which implies that $1/g(x)$ is convex. First, the second derivative is:

   $$\frac{\partial^2 \left(\frac{1}{g(x)}\right)}{\partial x^2} = \frac{2g'^2(x)}{g^3(x)} - \frac{g''(x)}{g^2(x)}$$

   Next, we argue that this function is always positive. First, since g is always positive, we know that $g^3(x) > 0$. Moreover, both $g'^2(x)$ and $g(x)$ are always greater than or equal to 0 (actually $g(x)$ is strictly greater than 0). Thus the first term is positive. Finally since g is concave, we know that $g''(x) \leq 0$, meaning that the second term is always always positive. Putting this together, we know that the second derivative of $1/g(x)$ is always positive, which implies that the function is convex.

   (b) Consider the following set: $S = \{(x, y) : x^2 \leq y < x^2 + 1\}$. All of the boundary points of this set lie on $y = x^2$, and they each have a supporting hyperplane. However, the set is clearly non-convex, as one can find two points such that the line passing between is not contained within the set. This only works because the set is not closed. If S were closed and had non-empty interior, then supporting hyperplanes would imply convexity.

2. (Subdifferentials)

   (a) The subdifferentials for $|z|$ are:

   $$\partial f(z) = \begin{cases} \text{sign}(z) : z \neq 0 \\ [-1, 1] : z = 0 \end{cases}$$

   Let's start by looking at the first case. If $z \neq 0$, then $|z|$ is just a line, with slope ± 1, depending on the sign of z. In particular, if $z > 0$, then $f(z) = z$, and the derivative is just 1. On the other hand, if $z < 0$, then $f(z) = -z$ and the derivative is -1. Thus for $z \neq 0$, the subdifferential is just the sign of z.

If $z = 0$, then using the definition of subdifferential, we look for y such that $f(z) \geq f(0) + yz$ for all z. First, we know that $f(0) = 0$. Next, if z is negative, then $f(z) = -z$, so we look for y such that $-z \geq yz$. Since z is negative, this holds for all $y \geq -1$. Looking at z that are positive, we'll get that $z \geq yz$, which holds for all $y \leq 1$. Thus we need that $-1 \leq y \leq 1$, and we get that the subdifferential at this point is $[-1, 1]$.

(b) First, we know that $||z||_1 = \sum_{i=1}^{n} |z_i|$. Using the definition of subdifferentials, we are looking for y such that $\sum_{i=1}^{n} |z_i| - |x_i| \geq y^T(z - x)$ for all z. If all of the x_is are non-zero, then it's easy to see that $y_i = sign(x_i)$; otherwise it's easy to violate the inequality (also this follows from the previous question). Similarly, if some $x_i = 0$, then we can have $y_i \in [-1, 1]$ without violating the inequality. Thus we see that the subdifferential is:

$$\partial f(z) = \left\{ y \in \mathbb{R}^n \,\middle|\, \begin{array}{r} y_i = \text{sign}(z_i) : z_i \neq 0 \\ y_i \in [-1, 1] : z_i = 0 \end{array} \right.$$

(c) We can write $f(z) = \frac{1}{2}||z - y||_2^2 + \lambda||z||_1$, as $\frac{1}{2}\sum_{i=1}^{n}(z_i - y_i)^2 + \lambda \sum_{i=1}^{n} |z_i|$. Additionally, we already know the subdifferential of the second part, and the subdifferential of the first part is just $z - y$. Putting these together, we get:

$$\partial f(z) = \left\{ x \,\middle|\, \begin{array}{l} x_i = z_i - y_i + \lambda \text{sign}(z_i) : z_i \neq 0 \\ x_i \in [-y_i - \lambda, -y_1 + \lambda] : z_i = 0 \end{array} \right.$$

(d) Finally, we search for z such that $0 \in \partial f(z)$. Setting $x_i = 0$ in each expression above gives:

$$\begin{aligned} z_i^* &= y_i - \lambda \text{sign}(z_i^*) \text{ if } z_i^* \neq 0 \\ y_i &\in [-\lambda, \lambda] \text{ if } z_i^* = 0 \end{aligned}$$

Next, noticing that the first expression translates into $z_i^* = y_i - \lambda$ if $z_i^* \geq 0$ and $z_i^* = y_i + \lambda$ if $z_i^* < 0$, we can then write:

$$\begin{aligned} z_i^* &= y_i - \lambda \text{ if } y_i > \lambda \\ z_i^* &= y_i + \lambda \text{ if } y_i < \lambda \\ z_i^* &= 0 \text{ if } y_i \in [-\lambda, \lambda] \end{aligned}$$

Which is exactly the expression we have for $z^*(i)$.

3. (Optimization)

(a) It's easy to see that the primal is a convex problem because (a) the objective is convex, (b) the inequality constraints are all convex (in fact they are affine), and (c) the equality constraints are affine.

First we'll show that $x_i \log x_i$ is convex. The first derivative is $\log x_i + 1$ and the second derivative of $x_i \log x_i$ is just $1/x_i$, which is positive for all $x_i > 0$ (i.e. all $x_i \in dom x_i \log x_i$). Thus each of the terms in the sum is convex, and consequently the sum is convex, by the non-negative weighted sum property.

It's easy to see that $Ax \leq b$ is affine since it is just a linear system of equations. Similarly $\mathbb{1}^T x = 1$ is also affine. Putting everything together, we see that the problem is convex.

(b) The KKT conditions are:

$$\begin{aligned} \lambda^* &\geq 0 && (1)\\ Ax^* &\leq b && (2)\\ \mathbb{1}^T x^* &= 1 && (3)\\ \lambda^*(Ax^* - b) &= 0 && (4)\\ \log x_i^* + 1 + \lambda^{*T} A^{(i)} + \mu^* &= 0 \ \forall i && (5) \end{aligned}$$

where $A^{(i)}$ denotes the ith column of A, i.e. we took all of the coefficients that we multiply by x_i, which is exactly the ith column.

(c) Suppose we are given λ^*. Then looking at Equation 5, we can write:

$$x_i^* = \exp\{-(\mu^* + 1 + \lambda^{*T} A^{(i)})\}$$

Then using Equation 3, we know that $\sum_{i=1}^n x_i^* = 1$, so we can normalize the x_i^*s (effectively solving for μ^*):

$$x_i^* = \frac{\exp\{-(\mu^* + 1 + \lambda^{*T} A^{(i)}\}}{\sum_{j=1}^n \exp\{-(\mu^* + 1 + \lambda^{*T} A^{(j)}\}} = \frac{\exp\{-\lambda^{*T} A^{(i)}\}}{\sum_{j=1}^n \exp\{-\lambda^{*T} A^{(j)}\}}$$

And these are the values of x^* in terms of just λ^*.

2 Density Estimation

(a) First we consider the bias.

$$\begin{aligned} \mathbb{E}(\hat{p}_h(x)) - p(x) &= \sum_{j=1}^N \mathbb{E}\left(\frac{\hat{\pi}_j}{h^d} \mathbb{1}(x \in B_j)\right) - p(x)\\ &= \sum_{j=1}^N \frac{1}{h^d n} \sum_{i=1}^n \mathbb{E}\left(\mathbb{1}(X_i \in B_j)\mathbb{1}(x \in B_j)\right) - p(x)\\ &= \sum_{j=1}^N \frac{1}{h^d} p(B_j) \mathbb{1}(x \in B_j) - p(x)\\ &= \sum_{j=1}^N \left(\frac{1}{h^d} p(B_j) - p(x)\right) \mathbb{1}(x \in B_j) \end{aligned}$$

Where $p(B_j) = \int_{B_j} p(x)dx$. Next, using the restrictions of $\mathcal{P}$ and since $1/h^d p(B_j)$ is the average density on the cube B_j (it has volume h^d, we know that we can upper bound $|p(x) - 1/h^d p(B_j)|$ by L times maximum two-norm difference between x and anything in B_j (i.e. $|p(x) - 1/h^d p(B_j)| \leq Lh\sqrt{d}$. This gives us:

$$\leq \sum_{j=1}^N Lh\sqrt{d} \mathbb{1}(x \in B_j) = Lh\sqrt{d}$$

Which is the bias.

For the variance, we reorder the sums to get $\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}\frac{\mathbb{1}(X_i\in B_j)}{h^d}\mathbb{1}(x \in B_j)$, and define $Y_i \triangleq \sum_{j=1}^{N}\frac{\mathbb{1}(X_i\in B_j)}{h^d}\mathbb{1}(x \in B_j)$. Then:

$$\begin{aligned} Var(\hat{p}_h(x)) &= \frac{1}{n}Var(Y_i) = \frac{1}{n}\left(\mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2\right) \leq \frac{1}{n}\mathbb{E}(Y_i^2) \\ \mathbb{E}(Y_i^2) &= \mathbb{E}\left(\sum_{j=1}^{N}\frac{\mathbb{1}(X_i \in B_j)}{h^{2d}}\mathbb{1}(x \in B_j)\right) \end{aligned}$$

Because all of the cross-terms cancel out (i.e. $\mathbb{1}(X_i \in B_j)\mathbb{1}(X_i \in B_k) = 0$ whenever $j \neq k$). Reducing this expression we get:

$$\mathbb{E}(Y_i^2) = \sum_{j=1}^{N}\frac{p(B_j)}{h^{2d}}\mathbb{1}(x \in B_j)$$

Plugging back into the expression for variance, we get:

$$Var(\hat{p}_h(x)) \leq \frac{1}{nh^{2d}}\sum_{j=1}^{N}p(B_j)\mathbb{1}(x \in B_j) \leq \frac{1}{nh^{2d}}$$

Thus the risk is upper bounded by $L^2h^2d + \frac{1}{nh^{2d}}$.

(b) We simply take the derivative (w.r.t h) of the expression for the upper bound for the risk, and solve for h. This gives the value of h in terms of n that minimizes this upper bound. The derivative is:

$$\begin{aligned} 2L^2dh - \frac{2d}{nh^{2d+1}} &= 0 \\ 2ndL^2h^{2d+2} - 2d &= 0 \\ h_n &= \left(\frac{1}{nL^2}\right)^{\frac{1}{2d+2}} \end{aligned}$$

Plugging this back into the upper bound for risk, we get:

$$\begin{aligned} & L^2d\left(\frac{1}{nL^2}\right)^{\frac{1}{d+1}} + \frac{1}{n\left(\frac{1}{nL^2}\right)^{\frac{2d}{2d+2}}} \\ = & \frac{ndL^2\left(\frac{1}{nL^2}\right)^{\frac{2d+2}{2d+2}} + 1}{n\left(\frac{1}{nL^2}\right)^{\frac{2d}{2d+2}}} \\ = & \frac{L^{\frac{d}{d+1}}d + 1}{n^{1-\frac{2d}{2d+2}}} \\ = & \frac{L^{\frac{d}{d+1}}d + 1}{n^{\frac{1}{d+1}}} \end{aligned}$$

So the risk goes to 0 at the rate $n^{\frac{-1}{d+1}}$, (i.e. in the one dimensional case the rate of convergence is $1/\sqrt{n}$

(c) By triangle inequality:

$$|\hat{p}_h(x) - p(x)| \le |\hat{p}_h(x) - p_h(x)| + |p_h(x) - p(x)| \le |\hat{p}_h(x) - p_h(x)| + Lh\sqrt{d}$$

Since the second term is exactly the bias (i.e. $p_h(x) = \mathbb{E}(\hat{p}_h(x))$). The next term, we can bound using Bernstein's inequality, since $\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i$ where each Y_i (defined above) is bounded by $\frac{1}{h^d}$ and has $Var(Y_i) \le \frac{1}{h^{2d}}$.

$$\mathbb{P}(|\hat{p}_h(x) - p_h(x)| > \epsilon) \le 2\exp\left\{-\frac{n\epsilon^2}{2h^{-2d} + 2h^{-d}\epsilon/3}\right\} = 2\exp\left\{\frac{-n\epsilon^2 h^{2d}}{2(1 + h^d\epsilon/3)}\right\}$$

Putting the triangle inequality expression together with this bound, we get:

$$\mathbb{P}(|\hat{p}_h(x) - p(x)| \ge \epsilon + Lh\sqrt{d}) \le 2\exp\left\{\frac{-n\epsilon^2 h^{2d}}{2(1 + h^d\epsilon/3)}\right\}$$

(d) First, using the triangle inequality, we'll get two terms that we need to bound:

$$\int |p(x) - \hat{p}_h(x)|dx \le \int |p(x) - p_h(x)|dx + \int |p_h(x) - \hat{p}_h(x)|dx$$

The first term is easy to bound. We break up the integral into each cube, and then recognize that $p_h(x) = p(B_j)/h^d$ on each cube. Applying the bound on the bias in Part (a) to every cube, we get a nice bound:

$$\int |p(x) - p_h(x)|dx = \sum_{j=1}^{n} \int_{B_j} |p(x) - p_h(x)|dx \le \sum_{j=1}^{N} Lh\sqrt{d} = NLh\sqrt{d}$$

For the second term, we'll use the multinomial inequality. First, we again break up the integral into the cubes, then recognize that both $\hat{p}_h$ and p_h are constant over each cube. This gives us:

$$\int |p_h(x) - \hat{p}_h(x)|dx = \sum_{j=1}^{n} h^d|\hat{p}_h - p_h|$$

Then, using that $h^d p_h = \mathbb{E}(h^d\hat{p}_h)$, and that this looks like a multinomial with n draws, we can directly apply the concentration inequality:

$$\mathbb{P}\left(\sum_{j=1}^{N} |h^d\hat{p}_h - h^d p_h| \ge n\epsilon\right) \le 3\exp\left\{\frac{-n\epsilon^2}{25}\right\}$$

Putting this together with the other half of the triangle inequality we'll get:

$$\mathbb{P}\left(\int |\hat{p}_h(x) - p(x)|dx) \ge n\epsilon + NLh\sqrt{d}\right) \le 3\exp\left\{\frac{-n\epsilon^2}{25}\right\}$$