Optimization-Based Data Analysis

Recitation 12

Remind Overton's class

1. What does NP-Hard mean?

   Solution. P denotes the class of decision problems that are solvable in polynomial time. NP (non-deterministic polynomial) denotes the class of problems for which solutions are checkable in polynomial time (equivalently, it denotes problems solvable on a non-deterministic Turing machine in polynomial time). A decision problem is NP-Hard if solving it in polynomial time implies a solution to all NP problems in polynomial time. More precisely, D is NP-Hard if every problem in NP has a polynomial time reduction to D. If $P \neq NP$ then NP-Hard problems are not solvable in polynomial time.

2. True or False: If S is a non-empty convex subset of $\mathbb{R}^n$ and $\vec{x} \in \mathbb{R}^n$ then there is a unique projection $\vec{y}$ of $\vec{x}$ onto S defined by

   $$\vec{y} = \arg\min_{\vec{y} \in S} \|\vec{y} - \vec{x}\|_2.$$

   Solution. False, the set has to be closed. For example, no projection of the point $(3, 3)$ onto the set $\{(x, y) \mid x^2 + y^2 < 1\}$.

3. Prove that $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$.

   Solution. Note that

   $$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots 1}{k!} \leq \frac{n^k}{k!}.$$

   To finish we must show that $\log(k!) \geq k \log(k/e)$. Note that

   $$\log(k!) = \sum_{j=2}^{k} \log(j) \geq \int_1^k \log(t)\, dt = k \log(k) - k + 1 \geq k \log(k/e),$$

   using Riemann sums since $\log(t)$ is monotone.

4. Consider the problem

   $$\begin{array}{ll} \text{minimize} & x^2 + y^2 \\ \text{subject to} & y = x^2 - 4. \end{array}$$

   (a) Is this a convex optimization problem?

   (b) What are the contour lines of $f_0(x, y) = e^{x^2+y}$?

   (c) What is the solution? Is it unique?

Solution.

(a) No, the feasible set isn't convex.

(b) Concentric circles centered at the origin.

(c) At the solution the gradient of the objective must be orthogonal to the feasible set. Note that

$$\nabla f_0(x, y) = (2x, 2y)$$

thus we need $(2x, 2y) = \alpha(-2x, 1)$ for some $\alpha \in \mathbb{R}$. Solving gives $x = 0$ and $y = \pm 4$ or $\alpha = -1$, $y = -1/2$, and $x = \pm\sqrt{3.5}$. The minimum is the latter, and it is not unique. This could have also been solved by determining where

$$\nabla_{x,y}[f_0(x, y) + \alpha(y - x^2 + 4)] = 0.$$

5. Let $A \in \mathbb{R}^{m\times n}$ with $n > m$, and suppose that $\vec{y} \in \mathbb{R}^m$ is in the column space of A.

(a) What is the solution to

$$\begin{array}{ll} \text{minimize} & \|\vec{x}\|_2^2 \\ \text{subject to} & A\vec{x} = \vec{y}? \end{array}$$

(b) Suppose we minimize $\frac{1}{2}\|A\vec{x} - \vec{y}\|_2^2$ over $\vec{x}$ using gradient descent started from the origin, and it converges to $\vec{x}^*$. Which of the infinitely many possible solutions will we obtain? What if we didn't start at the origin?

Solution.

(a) $\vec{x} = A^T(AA^T)^{-1}\vec{y}$, the pseudoinverse of A applied to $\vec{y}$. To see this must be the solution, note that it is an element of the row space of A. Any vector $\vec{x}$ can be written as $\vec{x} = \vec{x}_R + \vec{x}_N$ where $\vec{x}_R$ is in the row space of A and $\vec{x}_N$ is the null space of A. Then

$$\|\vec{x}\|_2^2 = \|\vec{x}_R\|_2^2 + \|\vec{x}_N\|_2^2,$$

but $A\vec{x} = A\vec{x}_R$, so the solution must lie in the row space.

(b) Each gradient step has the form

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - s\nabla f(\vec{x}) = \vec{x}^{(k)} - sA^T(A\vec{x}^{(k)} - \vec{y}).$$

Thus the step is always in the row space. If we start from the origin, this implies $\vec{x}^*$ is in the row space, and is thus the minimum ℓ^2-norm solution. Otherwise, it will be the minimum ℓ^2-norm solution plus the orthgonal projection of the starting point onto the null space of A.

6. Let $f, g : \mathbb{R}^2 \to \mathbb{R}$ be differentiable with $g(x, y) = f(2x, 5y)$.

(a) How does the gradient of g relate to f?

(b) What is the data science implication of this fact?

Solution.

(a) $\nabla g(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix} \nabla f(2x, 5y)$. For example, suppose we start at $(0, 0)$ and $\nabla f(0, 0) = (1, 1)$. Then $\nabla g(x, y) = (2, 5)$. So if we use a step size of 1 we go to $(1, 1)$ for f and $(2, 5)$ for g. But the equivalent step should be to $(1/2, 1/5)$.

(b) The scale of each feature has a noticeable effect on gradient descent. This isn't dealt with using backtracking line search (note that Newton's method doesn't have this issue).

7. A general linear program can be put in the following form:

$$\begin{array}{ll} \text{minimize}_{\vec{x}} & \vec{c}^T \vec{x} \\ \text{subject to} & A\vec{x} \succeq \vec{b}. \end{array}$$

(a) What is the associated Lagrangian?

(b) What is the dual problem?

(c) Fix $\vec{y}$ with $A\vec{y} \succeq \vec{b}$ and let $S = \{i \mid (A\vec{y})_i = \vec{b}_i\}$. Show that if there is a dual feasible $\vec{\lambda}$ with $\vec{\lambda}_j = 0$ for $j \notin S$ then $\vec{y}$ solves the primal problem.

Solution.

(a) $L(\vec{x}, \vec{\lambda}) = \vec{c}^T \vec{x} + \vec{\lambda}^T (\vec{b} - A\vec{x})$

(b) Define $g(\mu, \vec{\lambda}) = \min_{\vec{x}} L(\vec{x}, \vec{\lambda})$. Grouping terms in L gives

$$L(\vec{x}, \vec{\lambda}) = (\vec{c}^T - \vec{\lambda}^T A)\vec{x} + \vec{\lambda}^T \vec{b}.$$

Thus $g(\vec{\lambda}) = \vec{\lambda}^T \vec{b}$ if $\vec{c} = A^T \vec{\lambda}$ or $-\infty$ otherwise. This gives the dual problem

$$\begin{array}{ll} \text{maximize} & \vec{\lambda}^T \vec{b} \\ \text{subject to} & \vec{c} = A^T \vec{\lambda}, \\ & \vec{\lambda} \succeq 0. \end{array}$$

(c) Note that

$$\vec{c}^T \vec{y} = \vec{\lambda}^T A\vec{y} \succeq \vec{\lambda}^T \vec{b}$$

since $A\vec{y} \succeq \vec{b}$ and $\vec{\lambda} \succeq 0$. But by assumption $\vec{\lambda}^T A\vec{y} = \vec{\lambda}^T \vec{b}$ so by weak duality $\vec{y}$ is optimal for the primal.