Discussion 12

1. Newton's Method for Quadratic Functions

Give a symmetric positive definite matrix $Q \in \mathbb{S}^n_{++}$ and $b \in \mathbb{R}^n$, consider minimizing

$$f(x) = \frac{1}{2}\vec{x}^\top Q\vec{x} - \vec{b}^\top \vec{x} \tag{1}$$

Let $\vec{x}^\star$ denote the point at which $f(\vec{x})$ is minimized, and define $\mathcal{B}(\vec{x}^\star)$ as the ball centered at $\vec{x}^\star$ with unit ℓ_2 norm:

$$\mathcal{B}(\vec{x}^\star) = \{\vec{x} \in \mathbb{R}^n : \|\vec{x} - \vec{x}^\star\|_2 \leq 1\} \tag{2}$$

Assume we use Newton's method to minimize f:

$$\vec{x}_{k+1} = \vec{x}_k - (\nabla^2 f(\vec{x}_k))^{-1}\nabla f(\vec{x}_k) \tag{3}$$

where the initial point is $\vec{x}_0 \in \mathcal{B}(\vec{x}^\star)$.

For any $k \in \mathbb{N}$, find

$$\max_{\vec{x}_0 \in \mathcal{B}(\vec{x}^\star)} \|\vec{x}_k - \vec{x}^\star\|_2. \tag{4}$$

Solution: For all $\vec{x}$, the Hessian is $\nabla^2 f(\vec{x}) = Q$. Therefore, the update rule is

$$\vec{x}_{k+1} = \vec{x}_k - Q^{-1}(Q\vec{x}_k - \vec{b}) = Q^{-1}\vec{b} = \vec{x}^\star \tag{5}$$

So the optimal value of $\max_{\vec{x}_0 \in \mathcal{B}(\vec{x}^\star)} \|\vec{x}_k - \vec{x}^\star\|_2$ is 1 for $k = 0$ and 0 for $k \geq 1$

2. Generalized Linear Models

A wide class of machine learning models (e.g. classification and regression) can be modelled in a common framework called generalised linear models (GLMs). In this problem, we'll talk about exponential families, generalized linear models and use Newton's method to perform maximum likelihood estimation (MLEs). Consider a special class of probability distributions known as exponential families whose density is of the form

$$f(\vec{y}; \vec{\theta}) = e^{\vec{y}^\top \vec{\theta} - b(\vec{\theta})} f_0(\vec{y}) \tag{6}$$

where $\vec{y}, \vec{\theta} \in \mathbb{R}^n$ and $b(\vec{\theta}) = \log\left(\int_{\mathbb{R}^n} e^{\vec{y}^\top \vec{\theta}} f_0(\vec{y}) d\vec{y}\right)$ is the normalizing constant which ensures f is a probability distribution over $\vec{y}$.

(a) Show that $b(\vec{\theta})$ is a convex function.

Solution: In a previous homework, you proved that the following function $g(x_1, \ldots, x_n) = \log\left(\sum_{i=1}^n e^{x_i}\right)$ is convex. Certainly this implies that $g(A\vec{x} + \vec{c})$ is also convex for any $A, \vec{c}$. This can be seen as taking an A and $\vec{c}$ which are infinite dimensional $(n \to \infty)$ matrix and vector respectively as then the sum can be treated as an integral (recall Riemann sum). These are indexed by different values of $\vec{y}$ taking $A(\vec{y}) = \vec{y}^\top$ and $c(\vec{y}) = \log(f_0(\vec{y}))$. **We won't expect you to do infinite summations like this for the final but provide it here to give an example of such proofs.**

For more details on the above proof, recall the Riemann sum interpretation of integrals: $\int_a^b f(x)dx = \lim_{n\to\infty}\sum_{i=1}^n f(x_i)(x_i - x_{i-1})$ by picking n points $a = x_0 < x_1 < \ldots < x_n = b$ and making the distance between the points tending to zero by taking infinitely many points in this interval. This idea also works when the integral limits a and b are $-\infty$ and ∞. We do the proof for dimension $n = 1$ and the proof for higher dimensions is exactly similar. Now we see,

$$\log\left(\int_{-\infty}^{\infty} e^{y\theta}dy\right) = \log\left(\lim_{n\to\infty}\sum_{i=1}^{n} e^{y_i\theta}(y_i - y_{i-1})\right) \tag{7}$$

$$= \log\left(\lim_{n\to\infty}\sum_{i=1}^{n} e^{y_i\theta+\log(y_i-y_{i-1})}\right) \tag{8}$$

$$= \lim_{n\to\infty}\log\left(\sum_{i=1}^{n} e^{y_i\theta+\log(y_i-y_{i-1})}\right) \tag{9}$$

Now if $g(x_1,\ldots,x_n) = \log\left(\sum_{i=1}^n e^{x_i}\right)$ is convex, so is any affine transformation and so the above is convex. Convexity is preserved when taking limits so we get that $b(\vec{\theta})$ is convex.

(b) We model $\vec{\theta} = X\vec{\beta}$ where $X \in \mathbb{R}^{n\times d}$ is the data matrix. Under this parameterization of $\vec{\theta}$, the exponential family is called a generalized linear model. Prove that $b(X\vec{\beta})$ is convex in $\vec{\beta}$.

Solution: If $b(\vec{\theta})$ is convex, so is any linear transformation so $b(X\vec{\beta})$ is convex in $\vec{\beta}$.

(c) For a given exponential family/GLM model, MLE estimation for a data matrix X and corresponding output variables $\vec{y} \in \mathbb{R}^n$ corresponds to solving the following maximization problem:

$$\max_{\vec{\beta}} f(\vec{y}; X\vec{\beta}) \tag{10}$$

Prove that this maximization problem is equivalent to

$$\min_{\vec{\beta}} g(\vec{\beta}) \coloneqq -\vec{y}^\top X\vec{\beta} + b(X\vec{\beta}) \tag{11}$$

Show that this is a convex optimization problem. Which choice of $b(\cdot)$ recovers linear regression?

Solution: Consider the negative logarithm of $f(\vec{y}; X\vec{\beta})$

$$-\log\left(e^{\vec{y}^\top X\vec{\beta} - b(X\vec{\beta})} f_0(\vec{y})\right) = -\vec{y}^\top X\vec{\beta} + b(X\vec{\beta}) - \log(f_0(\vec{y})) \tag{12}$$

Since the last term doesn't depend on $\vec{\beta}$ we can ignore it. Since $\log(\cdot)$ is an increasing function, maximizing a function is same as maximizing logarithm of the function and hence equivalent to minimizing negative log of the function and so we are done. We recover linear regression by taking $b(\vec{x}) = \frac{1}{2}\|\vec{x}\|_2^2$. An algorithm for classification known as logistic regression can be recovered by taking $b(\vec{x}) = \sum_{i=1}^n \log(1 + e^{x_i})$.

(d) For the above convex minimization problem, find the undamped Newton's method (with step size 1) update. This update also goes by the name *iteratively reweighted least squares* (IRLS). Can you tell why? (For any iterate $\vec{\beta}$ and the Newton update on $\vec{\beta}$ denoted by $\vec{\beta}_+$, what optimization problem is $\vec{\beta}_+$ the optimum of?)

Solution:

$$\nabla_{\vec{\beta}} g(\vec{\beta}) = -X^\top\vec{y} + X^\top\nabla b(X\vec{\beta}) \tag{13}$$

$$\nabla^2_{\vec{\beta}} g(\vec{\beta}) = X^\top \nabla^2 b(X\vec{\beta}) X \tag{14}$$

So the Newton update looks like

$$\vec{\beta}_+ = \vec{\beta} - (X^\top H_{\vec{\beta}} X)^{-1}(X^\top \nabla_\beta b(\vec{\beta}) - X^\top \vec{y}) \tag{15}$$

$$= (X^\top H_{\vec{\beta}} X)^{-1} X^\top H_{\vec{\beta}} \vec{\alpha}_{\vec{\beta}} \tag{16}$$

where $H_{\vec{\beta}} = \nabla^2_{\vec{\beta}} b(X\vec{\beta})$ and $\vec{\alpha}_{\vec{\beta}} = X\vec{\beta} + H^{-1}_{\vec{\beta}}(\vec{y} - \nabla_{\vec{\beta}} b(X\vec{\beta}))$ You can verify that $\vec{\beta}_+$ is the solution to the following *weighted* least squares problem:

$$\min_{\vec{\theta}} \frac{1}{2} \|H^{1/2}_{\vec{\beta}}(X\vec{\theta} - \vec{\alpha}_{\vec{\beta}})\|_2^2 \tag{17}$$

so at each step we are solving a least squares like problem where we *reweigh* the observations by a PSD matrix $H^{1/2}_{\vec{\beta}}$ and change the output vector to $\vec{\alpha}_{\vec{\beta}}$. Note that both $H_{\vec{\beta}}$ and $\vec{\alpha}_{\vec{\beta}}$ depend on the current iterate $\vec{\beta}$. This is why this problem sometimes is a special case of a procedure known as *iteratively reweighted least squares*(IRLS)

3. Robust Linear Programming

In this problem we will consider a version of linear programming under uncertainty.

Consider vector $\vec{x} \in \mathbb{R}^n$. Recall from the previous discussion that $\vec{x}^\top \vec{y} \leq \|\vec{x}\|_1$ for all $\vec{y}$ such that $\|\vec{y}\|_\infty \leq 1$. Further this inequality is tight, since it holds with equality for $\vec{y} = \text{sgn}(\vec{x})$.

Let us focus now on a LP in standard form:

$$\min_{\vec{x}} \quad \vec{c}^\top \vec{x} \tag{18}$$

$$\text{s.t.} \quad \vec{a}_i^\top \vec{x} \leq b_i, \quad i = 1, ..., m. \tag{19}$$

Consider the set of linear inequalities in (19). Suppose you don't know the vectors $\vec{a}_i$ exactly. Instead you are given nominal values $\hat{\vec{a}}_i$, and you know that the actual vectors satisfy $\left\|\vec{a}_i - \hat{\vec{a}}_i\right\|_\infty \leq \rho$ for a given $\rho > 0$. In other words, the actual components a_{ij} can be anywhere in the intervals $[\hat{a}_{ij} - \rho, \hat{a}_{ij} + \rho]$. Or equivalently, each vector $\vec{a}_i$ can lie anywhere in a hypercube with corners $\hat{\vec{a}}_i + \vec{v}$ where $\vec{v} \in \{-\rho, \rho\}^n$. We desire that the set of inequalities that constrain problem (19) be satisfied for all possible values of $\vec{a}_i$; i.e., we replace these with the constraints

$$\vec{a}_i^\top \vec{x} \leq b_i \;\; \forall \vec{a}_i \in \{\hat{\vec{a}}_i + \vec{v} \mid \|\vec{v}\|_\infty \leq \rho\} \;\; i = 1, ..., m. \tag{20}$$

Note that the above defines an *infinite* number of constraints (of the form $\hat{\vec{a}}_i^\top \vec{x} + \vec{v}^\top \vec{x} \leq b_i, \quad \forall \vec{v}$ satisfying $\|\vec{v}\|_\infty \leq \rho, i = 1, 2, \ldots, m$).

(a) Argue why for our LP we can replace the infinite set of constraints as above to a finite set of $2^n m$ constraints of the form,

$$\hat{\vec{a}}_i^\top \vec{x} + \vec{v}^\top \vec{x} \leq b_i \;\; \forall \vec{v} \in \{-\rho, \rho\}^n \;\; i = 1, ..., m. \tag{21}$$

HINT: What do you know about the optimal solutions of LPs?

Solution: We know from LP theory that the maximum of an affine function on a bounded convex set must occur at an extreme point of the set. Thus we only need to consider the maximum value the LHS of Equation (20) could take. Consequently for each i we can go from infinite constraints to 2^n constraints

since the set $\|\vec{v}\|_\infty \leq \rho$ set has 2^n extreme points (corners of the n-dimensional hypercube). Doing this we obtain the constraint set,

$$\vec{\hat{a}}_i^\top \vec{x} + \vec{v}^\top \vec{x} \leq b_i \;\; \forall \vec{v} \in \{-\rho, \rho\}^n \;\; i = 1, ..., m. \tag{22}$$

(b) Use result from part (a) to show that the constraint set in Equation (20) is in fact equivalent to the much more compact set of m nonlinear inequalities

$$\vec{\hat{a}}_i^\top \vec{x} + \rho \|\vec{x}\|_1 \leq b_i, \;\; i = 1, ..., m. \tag{23}$$

Solution: From Equation (20) we have the constraints,

$$\vec{a}_i^\top \vec{x} \leq b_i \;\; \forall \vec{a}_i \in \{\vec{\hat{a}}_i + \vec{v} \mid \|\vec{v}\|_\infty \leq \rho\} \;\; i = 1, ..., m. \tag{24}$$

This inequality can be written as:

$$\vec{\hat{a}}_i^\top \vec{x} + \vec{v}^\top \vec{x} \leq b_i \;\; \forall \vec{v} \in \{v \mid \|\vec{v}\|_\infty \leq \rho\} \;\; i = 1, ..., m. \tag{25}$$

or equivalently as,

$$\vec{\hat{a}}_i^\top \vec{x} + \rho \vec{v}^\top \vec{x} \leq b_i \;\; \forall \vec{v} \in \{\vec{v} \mid \|\vec{v}\|_\infty \leq 1\} \;\; i = 1, ..., m. \tag{26}$$

This is equivalent to ,

$$\vec{\hat{a}}_i^\top \vec{x} + \rho \max_{\|\vec{v}\|_\infty \leq 1} \vec{v}^\top \vec{x} \leq b_i \;\; \forall i = 1, ..., m. \tag{27}$$

because if the inequality is satisfied when the LHS is maximized over $\vec{v}$ such that $\|\vec{v}\|_\infty \leq 1$ then it is satisfied for all $\vec{v}$ such that $\|\vec{v}\|_\infty \leq 1$. From part (a), we have $\max_{\|\vec{y}\|_\infty \leq 1} \vec{x}^\top \vec{y} = \|\vec{x}\|_1$, which gives us the equivalent constraint set:

$$\vec{\hat{a}}_i^\top \vec{x} + \rho \|\vec{x}\|_1 \leq b_i, \;\; i = 1, ..., m. \tag{28}$$

We now would like to formulate the LP with uncertainty introduced. We are therefore interested in situations where the vectors $\vec{a}_i$ are uncertain, but satisfy bounds $\left\|\vec{a}_i - \vec{\hat{a}}_i\right\|_\infty \leq \rho$ for given $\vec{\hat{a}}_i$ and ρ. We want to minimize $\vec{c}^\top \vec{x}$ subject to the constraint that the inequalities $\vec{a}_i^\top \vec{x} \leq b_i$ are satisfied for *all* possible values of $\vec{a}_i$.
We call this a *robust LP* :

$$\min_{\vec{x}} \quad \vec{c}^\top \vec{x} \tag{29}$$

$$\text{s.t.} \quad \vec{a}_i^\top \vec{x} \leq b_i, \;\; \forall \vec{a}_i \in \{\vec{\hat{a}}_i + \vec{v} \mid \|\vec{v}\|_\infty \leq \rho\} \;\; i = 1, ..., m. \tag{30}$$

(c) Using the result from part (c), express the above optimization problem as an LP.

Solution: From part (c), We can rewrite the problem as

$$\min_{\vec{x}} \quad \vec{c}^\top \vec{x} \tag{31}$$

$$\text{s.t.} \quad \vec{a}_i^\top \vec{x} + \rho \left\| \vec{x} \right\|_1 \leq b_i, \quad i = 1, ..., m. \tag{32}$$

We can express this optimization problem as an LP by introducing variables t_i:

$$\min_{\vec{x}, \vec{t}} \quad \vec{c}^\top \vec{x} \tag{33}$$

$$\text{s.t.} \quad \vec{a}_i^\top \vec{x} + \rho \sum_i t_i \leq b_i, \quad i = 1, ..., m. \tag{34}$$

$$x_i \leq t_i \quad i = 1, ..., m. \tag{35}$$

$$-x_i \leq t_i \quad i = 1, ..., m. \tag{36}$$