Statistical Models & Computing Methods

Lecture 2: Optimization

Cheng Zhang
School of Mathematical Sciences, Peking University

September 13, 2022

- Consider the following least square problem

$$\text{minimize} \quad L(\beta) = \frac{1}{2}\|Y - X\beta\|^2$$

- Note that this is a quadratic problem, which can be solved by setting the gradient to zero

$$\nabla_\beta L(\beta) = -X^T(Y - X\hat{\beta}) = 0$$
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

given that the Hessian is positive definite:

$$\nabla^2 L(\beta) = X^T X \succ 0$$

which is true iff X has independent columns.

Regularized Regression Models

- In practice, we would like to solve the least square problems with some constraints on the parameters to control the complexity of the resulting model
- One common approach is to use Bridge regression models (Frank and Friedman, 1993)

$$\begin{aligned} \text{minimize} \quad & L(\beta) = \frac{1}{2}\|Y - X\beta\|^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} |\beta_j|^\gamma \leq s \end{aligned}$$

- Two important special cases are ridge regression (Hoerl and Kennard, 1970) $\gamma = 2$ and Lasso (Tibshirani, 1996) $\gamma = 1$

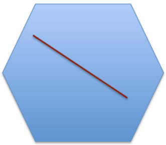
- In general, optimization problems take the following form:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

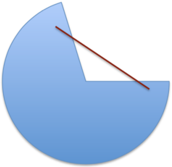
- We are mostly interested in **convex** optimization problems, where the objective function $f_0(x)$, the inequality constraints $f_i(x)$ and the equality constraints $h_j(x)$ are all convex functions.

- A set C is *convex* if the line segment between any two points in C also lies in C, i.e.,

$$\theta x_1 + (1-\theta)x_2 \in C, \quad \forall x_1, x_2 \in C,\ 0 \le \theta \le 1$$



Convex Set

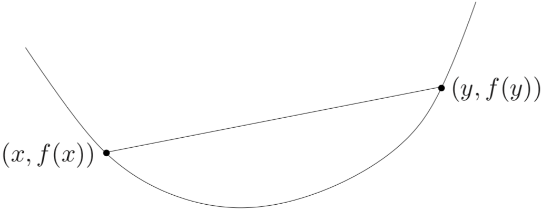


Non-convex Set

- If C is a convex set in $\mathbb{R}^n$ and $f(x) : \mathbb{R}^n \to \mathbb{R}^n$ is an affine function, then $f(C)$, i.e., the image of C is also a convex set.

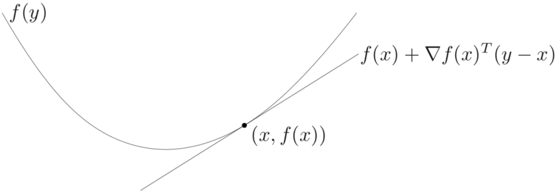
- A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if its domain D_f is a convex set, and $\forall x, y \in D_f$ and $0 \leq \theta \leq 1$

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$



- For example, many norms are convex functions

$$\|x\|_p = (\sum_i |x_i|^p)^{1/p}, \quad p \geq 1$$



- First order conditions. Suppose f is differentiable, then f is convex iff D_f is convex and

$$f(y) \geq f(x) + \nabla f(x)^T(y-x), \quad \forall x, y \in D_f$$

Corollary: For convex function f,

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$$

- Second order conditions. $\nabla^2 f(x) \succeq 0, \ \forall x \in D_f$

Basic Terminology and Notations

- Optimal value $p^* = \inf\{f_0(x) | f_i(x) \leq 0,\ h_j(x) = 0\}$
- x is feasible if $x \in D = \bigcap_{i=0}^{m} D_{f_i} \cap \bigcap_{j=1}^{p} D_{h_j}$ and satisfies the constraints.
- A feasible x^* is optimal if $f(x^*) = p^*$
- Optimality criterion. Assuming f_0 is convex and differentiable, x is optimal iff

$$\nabla f_0(x)^T (y - x) \geq 0, \quad \forall \text{ feasible y}$$

Remark: for unconstrained problems, x is optimial iff

$$\nabla f_0(x) = 0$$

Local Optimality

x is locally optimal if for a given $R > 0$, it is optimal for

$$
\begin{array}{rl}
\text{minimize} & f_0(z) \\
\text{subject to} & f_i(z) \leq 0, \quad i = 1, \ldots, m \\
& h_j(z) = 0, \quad j = 1, \ldots, p \\
& \|z - x\| \leq R
\end{array}
$$

In convex optimization problems, any locally optimal point is also globally optimal.

The Lagrangian

- Consider a general optimization problem

$$
\begin{array}{rll}
\text{minimize} & f_0(x) & \\
\text{subject to} & f_i(x) \leq 0, & i = 1, \ldots, m \\
& h_j(x) = 0, & j = 1, \ldots, p
\end{array}
$$

- To take the constraints into account, we augment the objective function with a weighted sum of the constraints and define the **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ as

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$$

  where λ and ν are dual variables or *Lagrangian multipliers*.

- We define the Lagrangian dual function as follows

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu)$$

- The dual function is the pointwise infimum of a family of affine functions of (λ, ν), it is concave, even when the original problem is not convex.
- If $\lambda \geq 0$, for each feasible point $\tilde{x}$

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) \leq L(\tilde{x}, \lambda, \nu) \leq f_0(\tilde{x})$$

- Therefore, $g(\lambda, \nu)$ is a lower bound for the optimial value

$$g(\lambda, \nu) \leq p^*, \quad \forall \lambda \geq 0, \nu \in \mathbb{R}^p$$

The Lagrangian Dual Problem

- Finding the best lower bound leads to the Lagrangian dual problem

$$\text{maximize} \;\; g(\lambda, \nu), \quad \text{subject to} \;\; \lambda \geq 0$$

- The above problem is a convex optimization problem.
- We denote the optimal value as d^*, and call the corresponding solution (λ^*, ν^*) the dual optimal
- In contrast, the original problem is called the primal problem, whose solution x^* is called primal optimal

- d^* is the best lower bound for p^* that can be obtained from the Lagrangian dual function.
- Weak Duality
$$d^* \leq p^*$$
- The difference $p^* - d^*$ is called the *optimal dual gap*
- Strong Duality
$$d^* = p^*$$

- Strong duality doesn't hold in general, but if the primal is convex, it usually holds under some conditions called *constraint qualifications*
- A simple and well-known constraint qualification is Slater's condition: there exist an x in the relative interior of D such that

$$f_i(x) < 0, \quad i = 1, \ldots, m, \qquad Ax = b$$

Complementary Slackness

- Consider primal optmial x^* and dual optimal (λ^*, ν^*)
- If strong duality holds

$$
\begin{aligned}
f_0(x^*) &= g(\lambda^*, \nu^*) \\
&= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} v_j^* h_i(x) \right) \\
&\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} v_j^* h_i(x^*) \\
&\leq f_0(x^*).
\end{aligned}
$$

- Therefore, these are all equalities

Complementary Slackness

- Important conclusions:
  - x^* minimize $L(x, \lambda^*, \nu^*)$
  - $\lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m$
- The latter is called complementary slackness, which indicates

$$\begin{aligned} \lambda_i^* > 0 \quad &\Rightarrow \quad f_i(x^*) = 0 \\ f_i(x^*) < 0 \quad &\Rightarrow \quad \lambda_i^* = 0 \end{aligned}$$

- When the dual problem is easier to solve, we can find (λ^*, ν^*) and then minimize $L(x, \lambda^*, \nu^*)$. If the resulting solution is primal feasible, then it is primal optimal.

Entropy Maximization

- Consider the entropy maximization problem

$$\begin{aligned} \text{minimize} \quad & f_0(x) = \sum\nolimits_{i=1}^{n} x_i \log x_i \\ \text{subject to} \quad & -x_i \le 0, \quad i = 1, \ldots, n \\ & \sum\nolimits_{i=1}^{n} x_i = 1 \end{aligned}$$

- Lagrangian

$$L(x, \lambda, \nu) = \sum_{i=1}^{n} x_i \log x_i - \sum_{i=1}^{n} \lambda_i x_i + \nu\Big(\sum_{i=1}^{n} x_i - 1\Big)$$

- We minimize $L(x, \lambda, \mu)$ by setting $\frac{\partial L}{\partial x}$ to zero

$$\log \hat{x}_i + 1 - \lambda_i + \nu = 0 \Rightarrow \hat{x}_i = \exp(\lambda_i - \nu - 1)$$

- The dual function is

$$g(\lambda, \nu) = -\sum_{i=1}^{n} \exp(\lambda_i - \nu - 1) - \nu$$

- Dual:

$$\text{maximize} \ \ g(\lambda, \nu) = -\exp(-\nu - 1) \sum_{i=1}^{n} \exp(\lambda_i) - \nu, \quad \lambda \geq 0$$

- We find the dual optimal

$$\lambda_i^* = 0, \ \ i = 0, \ldots, n, \quad \nu^* = -1 + \log n$$

- We now minimize $L(x, \lambda^*, \nu^*)$

$$\log x_i^* + 1 - \lambda_i^* + \nu^* = 0 \quad \Rightarrow \quad x_i^* = \frac{1}{n}$$

- Therefore, the discrete probability distribution that has maximum entropy is the uniform distribution

Exercise

Show that $X \sim \mathcal{N}(\mu, \sigma^2)$ is the maximum entropy distribution such that $EX = \mu$ and $EX^2 = \mu^2 + \sigma^2$. How about fixing the first k moments at $EX^i = m_i,\ i = 1, \ldots, k$?

Karush-Kun-Tucker (KKT) conditions

- Suppose the functions $f_0, f_1, \ldots, f_m, h_1, \ldots, h_p$ are all differentiable; x^* and (λ^*, ν^*) are primal and dual optimal points with zero duality gap
- Since x^* minimize $L(x, \lambda^*, \nu^*)$, the gradient vanishes at x^*

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^{p} \nu_i^* \nabla h_j(x^*) = 0$$

- Additionally

$$\begin{aligned} f_i(x^*) &\leq 0, \quad i = 1, \ldots, m \\ h_j(x^*) &= 0, \quad j = 1, \ldots, p \\ \lambda_i^* &\geq 0, \quad i = 1, \ldots, m \\ \lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m \end{aligned}$$

- These are called Karush-Kuhn-Tucker (KKT) conditions

KKT conditions for convex problems

- When the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal with zero duality gap.
- Let $\tilde{x}, \tilde{\lambda}, \tilde{\nu}$ be any points that satisfy the KKT conditions, $\tilde{x}$ is primal feasible and minimizes $L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$

$$\begin{aligned} g(\tilde{\lambda}, \tilde{\nu}) &= L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \\ &= f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(\tilde{x}) + \sum_{j=1}^{p} \tilde{\nu}_j h_j(\tilde{x}) \\ &= f_0(\tilde{x}) \end{aligned}$$

- Therefore, for convex optimization problems with differentiable functions that satisfy Slater's condition, the KKT condtions are necessary and sufficient

- Consider the following problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^T P x + q^T x + r, \quad P \succeq 0 \\ \text{subject to} & Ax = b \end{array}$$

- KKT conditions:

$$\begin{aligned} Px^* + q + A^T \nu^* &= 0 \\ Ax^* &= b \end{aligned}$$

- To find x^*, v^*, we can solve the above system of linear equations

- We now focus on numerical solutions for unconstrained optimization problems

$$\text{minimize} \quad f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable

- Descent method. We can set up a sequence

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}, \quad t^{(k)} > 0$$

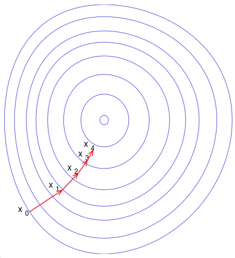
such that $f(x^{(k+1)}) < f(x^{(k)}), \quad k = 0, 1, \ldots,$

- $\Delta x^{(k)}$ is called the search direction; $t^{(k)}$ is called the step size or learning rate in machine learning.

A reasonable choice for the search direction is the negative gradient, which leads to gradient descent methods

$$x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x^{(k)}), \quad k = 0, 1, \ldots$$

- step size $t^{(k)}$ can be constant or determined by line search
- every iteration is cheap, does not require second derivatives



- First-order Taylor expansion

$$f(x+v) \approx f(x) + \nabla f(x)^T v$$

- v is a descent direction iff $\nabla f(x)^T v < 0$
- Negative gradient is the steepest descent direction with respect to the Euclidean norm.

$$\frac{-\nabla f(x)}{\|\nabla f(x)\|_2} = \arg\min_{v} \{\nabla f(x)^T v \mid \|v\|_2 = 1\}$$

- Consider the second-order Taylor expansion of f at x,

$$\begin{aligned} f(x+v) &\approx f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v \\ &\triangleq \tilde{f}(x) \end{aligned}$$

- We find the optimal direction v by minimizing $\tilde{f}(x)$ with respect to v

$$v = -[\nabla^2 f(x)]^{-1} \nabla f(x)$$

- If $\nabla^2 f(x) \succeq 0$ (e.g., convex functions)

$$\nabla f(x)^T v = -\nabla f(x)^T [\nabla^2 f(x)]^{-1} \nabla f(x) < 0$$

when $\nabla f(x) \neq 0$

- The search direction in Newton's method can also be viewed as a steepest descent direction, but with a different metric
- In general, given a positive definite matrix P, we can define a quadratic norm

$$\|v\|_P = (v^T P v)^{1/2}$$

- Similarly, we can show that $-P^{-1}\nabla f(x)$ is the steepest descent direction w.r.t. the quadratic norm $\|\cdot\|_P$

$$\text{minimize } \nabla f(x)^T v, \quad \text{subject to } \|v\|_P = 1$$

- When P is the Hessian $\nabla^2 f(x)$, we get Newton's method

Quasi-Newton Method

- Computing the Hessian and its inverse could be expensive, we can approximate it with another positive definite matrix $M \succ 0$ which is easier to use
- Update $M^{(k)}$ to learn about the curvature of f in the search direction and maintain a secant condition

$$\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) = M^{(k+1)}(x^{(k+1)} - x^{(k)})$$

- Rank-one update

$$\begin{aligned}
\Delta x^{(k)} &= x^{(k+1)} - x^{(k)} \\
y^{(k)} &= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \\
v^{(k)} &= y^{(k)} - M^{(k)} \Delta x^{(k)} \\
M^{(k+1)} &= M^{(k)} + \frac{v^{(k)} (v^{(k)})^T}{(v^{(k)})^T \Delta x^{(k)}}
\end{aligned}$$

- Easy to compute the inverse of matrices for low rank updates by **Sherman-Morrison-Woodbury formula**

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

  where $A \in \mathbb{R}^{n\times n}, U \in \mathbb{R}^{n\times d}, C \in \mathbb{R}^{d\times d}, V \in \mathbb{R}^{d\times n}$

- Another popular rank-two update method: the BFGS (Broyden-Fletcher-Goldfarb-Shanno) method

$$M^{(k+1)} = M^{(k)} + \frac{y^{(k)}(y^{(k)})^T}{(y^{(k)})^T\Delta x^{(k)}} - \frac{M^{(k)}\Delta x^{(k)}(M^{(k)}\Delta x^{(k)})^T}{(\Delta x^{(k)})^T M^{(k)}\Delta x^{(k)}}$$

Maximum Likelihood Estimation

- In the frequentist framework, we typically perform statistical inference by maximizing the log-likelihood $L(\theta)$, or equivalently minimizing negative log-likelihood, which is also known as the energy function
- Some notations we introduced before
  - Score function: $s(\theta) = \nabla_\theta L(\theta)$
  - Observed Fisher information: $J(\theta) = -\nabla_\theta^2 L(\theta)$
  - Fisher information: $\mathcal{I}(\theta) = \mathbb{E}(-\nabla_\theta^2 L(\theta))$
- Newton's method for MLE:

$$\theta^{(k+1)} = \theta^{(k)} + (J(\theta^{(k)}))^{-1} s(\theta^{(k)})$$

Fisher Scoring Algorithm

- If we use the Fisher information instead of the observed information, the resulting method is called the *Fisher scoring* algorithm

$$\theta^{(k+1)} = \theta^{(k)} + (\mathcal{I}(\theta^{(k)}))^{-1} s(\theta^{(k)})$$

- It seems that the Fisher scoring algorithm is less sensitive to the initial guess. On the other hand, the Newton's method tends to converge faster
- For exponential family models with natural parameters and generalized linear models (GLMs) with canonical links, the two methods are identical

- A generalized linear model (GLM) assumes a set of independent random variables $Y_1, \dots, Y_n$ that follow exponential family distributions of the same form

$$p(y_i|\theta_i) = \exp\left(y_i b(\theta_i) + c(\theta_i) + d(y_i)\right)$$

- The parameters θ_i are typically not of direct interest. Instead, we usually assume that the expectation of Y_i can be related to a vector of parameters β via a transformation (link function)

$$E(Y_i) = \mu_i, \quad g(\mu_i) = x_i^T \beta$$

where x_i is the observed covariates for y_i.

- Using the link function, we can now write the score function in terms of β
- Let $g(\mu_i) = \eta_i$, we can show that for jth parameter

$$s(\beta_j) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}\mathrm{ar}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

  where $\partial\mu_i/\partial\eta_i$ depends on the link function we choose
- It is also easy to show that the Fisher information matrix is

$$\begin{aligned}\mathcal{I}(\beta_j, \beta_k) &= \mathbb{E}(s(\beta_j)s(\beta_k)) \\ &= \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\mathbb{V}\mathrm{ar}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\end{aligned}$$

- Note that the Fisher information matrix can be written as

$$\mathcal{I}(\beta) = X^T W X$$

where W is the $n \times n$ diagonal matrix with elements

$$w_{ii} = \frac{1}{\mathbb{V}\mathrm{ar}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

- Rewriting Fisher scoring algorithm for updating β as

$$\mathcal{I}(\beta^{(k)})\beta^{(k+1)} = \mathcal{I}(\beta^{(k)})\beta^{(k)} + s(\beta^{(k)})$$

- After few simple steps, we have

$$X^T W^{(k)} X \beta^{(k+1)} = X^T W^{(k)} Z^{(k)}$$

where

$$z_i^{(k)} = \eta_i^{(k)} + (y_i - \mu_i^{(k)}) \frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}}$$

- Therefore, we can find the next estimate as follows

$$\beta^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} Z^{(k)}$$

- The above estimate is similar to the weighted least square estimate, except that the weights W and the response variable Z change from one iteration to another
- We iteratively estimate β until the algorithm converges

Example: Logistic Regression

- Recall that the Log-likelihood for logistic regression is

$$L(Y|p) = \sum_{i=1}^{n} y_i \log \frac{p_i}{1-p_i} + \log(1-p_i)$$

- The natural parameters are $\theta_i = \log \frac{p_i}{1-p_i}$. We use $g(x) = \log \frac{x}{1-x}$ as the link function, $\theta_i = g(p_i) = x_i^T \beta$
- We now write the log-likelihood as follows

$$L(\beta) = Y^T X\beta - \sum_{i=1}^{n} \log(1 + \exp(x_i^T \beta))$$

- The score function is

$$s(\beta) = X^T(Y-p), \quad p = \frac{1}{1+\exp(-X\beta)}$$

- The observed Fisher information matrix is

$$J(\beta) = X^T W X$$

where W is a diagonal matrix with elements

$$w_{ii} = p_i(1 - p_i)$$

- Note that $J(\beta)$ does not depend on Y, meaning that it is also the Fisher information matrix $\mathcal{I}(\beta) = J(\beta)$
- Newton's update

$$\beta^{(k+1)} = \beta^{(k)} + \left(X^T W^{(k)} X\right)^{-1} \left(X^T(Y - p^{(k)})\right)$$

Reference

- I. Frank and J. Friedman. A statistical view of some chemometrics regression tools (with discussion). Technometrics, 35, 109-148, (1993)
- A. Hoerl and R. Kennard. Ridge regression. In Encyclopedia of Statistical Sciences, 8, 129-136, 1988
- R. Tibshirani. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267-288. 1996
- S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, 2004