Introduction to Machine Learning
Answers to Exercise 1

Jingtao Min

March 7, 2022

1 Multivariate normal distribution

(a) Let d-dimensional random vector $\mathbf{X} = (X_1, ..., X_d)$ follow standard Gaussian distribution, i.e. $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$. Define $\mathbf{Y} = A\mathbf{X} + \boldsymbol{\mu}$ where $A \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$. Then the characteristic function:

$$\begin{aligned}\varphi_{\mathbf{Y}}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(i\mathbf{t}^T\mathbf{Y}\right)\right] = \mathbb{E}\left[\exp\left(i\mathbf{t}^T(A\mathbf{X}+\boldsymbol{\mu})\right)\right] = \exp\left(i\mathbf{t}^T\boldsymbol{\mu}\right)\mathbb{E}\left[\exp\left(i\mathbf{t}^TA\mathbf{X}\right)\right] \\ &= \exp\left(i\mathbf{t}^T\boldsymbol{\mu}\right)\mathbb{E}\left[\exp\left(i\left(A^T\mathbf{t}\right)^T\mathbf{X}\right)\right] = \exp\left(i\mathbf{t}^T\boldsymbol{\mu}\right)\exp\left[i\left(A^T\mathbf{t}\right)^T\mathbf{0} - \frac{1}{2}\left(A^T\mathbf{t}\right)^T\mathbf{I}\left(A^T\mathbf{t}\right)\right] \\ &= \exp\left(i\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^TAA^T\mathbf{t}\right) = \exp\left(i\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}\right)\end{aligned} \tag{1}$$

Therefore $\mathbf{Y} \sim \mathcal{N}_d(\boldsymbol{\mu}, AA^T)$.

(b) Let $B \in \mathbb{R}^{r\times d}$, apply the same procedure to $B\mathbf{Y}$:

$$\begin{aligned}\varphi_{B\mathbf{Y}}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(i\mathbf{t}^TB\mathbf{Y}\right)\right] = \mathbb{E}\left[\exp\left(i(B^T\mathbf{t})^T\mathbf{Y}\right)\right] \\ &= \exp\left(i(B^T\mathbf{t})^T\boldsymbol{\mu} - \frac{1}{2}(B^T\mathbf{t})^T(AA^T)(B^T\mathbf{t})\right) \\ &= \exp\left(i\mathbf{t}^TB\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^TBAA^TB^T\mathbf{t}\right)\end{aligned} \tag{2}$$

where $\mathbf{t} \in \mathbb{R}^r$. Therefore $B\mathbf{Y} \sim \mathcal{N}_r(B\boldsymbol{\mu}, BAA^TB^T)$.

(c) Let $\mathbf{X} = (X_1, X_2)$ be a bivariate normal random variable with mean $\mu = (1, 1)$ and covariance $\Sigma = \begin{pmatrix}3 & 1\\ 1 & 2\end{pmatrix}$. Note that the two random variables desired can be given by:

$$\begin{pmatrix}Y\\ Z\end{pmatrix} = \begin{pmatrix}X_1+X_2\\ X_1-X_2\end{pmatrix} = \begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}X_1\\ X_2\end{pmatrix} = \begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\mathbf{X} = A\mathbf{X} \tag{3}$$

Invoking the conclusion of the previous question, we know that (Y, Z) is also a Gaussian random vector:

$$\mathbf{X}' = \begin{pmatrix}Y\\ Z\end{pmatrix} \sim \mathcal{N}_2\left(A\boldsymbol{\mu}, A\Sigma A^T\right) = \mathcal{N}_2\left(\begin{pmatrix}2\\ 0\end{pmatrix}, \begin{pmatrix}7 & 1\\ 1 & 3\end{pmatrix}\right) = \mathcal{N}_2\left(\boldsymbol{\mu}', \Sigma'\right) \tag{4}$$

Recall that the density function is proportional to $\exp\{-\frac{1}{2}(\mathbf{X}'-\boldsymbol{\mu})^T(\Sigma')^{-1}(\mathbf{X}'-\boldsymbol{\mu})\}$, the conditional distribution can be obtained by setting $Z = 0$. Since $\mu_2' = \bar{Z} = 0$, the conditional distribution can be rewritten as:

$$\begin{aligned}p_{Y|Z}(y|0) = \frac{p_{Y,Z}(y,0)}{p_Z(0)} &\propto p_{Y,Z}(y,0) = \exp\left\{-\frac{1}{2}\begin{pmatrix}y-2\\ 0\end{pmatrix}^T\begin{pmatrix}7 & 1\\ 1 & 3\end{pmatrix}^{-1}\begin{pmatrix}y-2\\ 0\end{pmatrix}\right\} \\ &\propto \exp\left\{-\frac{(y-2)^2}{2\times\det(\Sigma)}/3\right\} = \exp\left\{-\frac{1}{2}\frac{(y-2)^2}{20/3}\right\}\end{aligned} \tag{5}$$

Therefore the conditional distribution $Y|_{Z=0} \sim \mathcal{N}_2\left(2, \frac{20}{3}\right)$.

2 Local vs. global optima

(a) A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if $\text{dom} f$ is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \text{dom} f$, and $0 \leq \theta \leq 1$, we have $f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$. It is strictly convex if strict inequality holds whenever $\mathbf{x} \neq \mathbf{y}$ and $0 < \theta < 1$.

(1) **The sum of one *strictly convex* function and *convex* functions is *strictly convex*.**

Proof:

Let g be a *strictly convex* function, h be a convex function, which share a convex domain Ω. Let $f = g + h$, so $\forall \mathbf{x} \neq \mathbf{y} \in \Omega$ and $0 < \theta < 1$, we have:

$$\begin{aligned} f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) &= g(\theta\mathbf{x} + (1-\theta)\mathbf{y}) + h(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \\ &< [\theta g(\mathbf{x}) + (1-\theta) g(\mathbf{y})] + [\theta h(\mathbf{x}) + (1-\theta) h(\mathbf{y})] \\ &= \theta\,[g(\mathbf{x}) + h(\mathbf{x})] + (1-\theta)\,[g(\mathbf{y}) + h(\mathbf{y})] \\ &= \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \quad \square \end{aligned} \tag{6}$$

(2) **Any local minimum of a convex function is also a global minimum.**

Proof:

Let f be a *convex* function. $\mathbf{x}_0$ is a local minimum, so that in its δ-neighbourhood, $f(\mathbf{x}) \geq f(\mathbf{x}_0)$. Suppose $\exists \mathbf{x}^* \in \Omega$ such that $f(\mathbf{x}^*) < f(\mathbf{x}_0)$. According to convexity, for a chosen $\theta = \delta / 2\,\|\mathbf{x}_0 - \mathbf{x}^*\|$,

$$f(\mathbf{x}') = f(\theta\mathbf{x}_* + (1-\theta)\mathbf{x}_0) \leq \theta f(\mathbf{x}^*) + (1-\theta) f(\mathbf{x}_0) < \theta f(\mathbf{x}_0) + (1-\theta) f(\mathbf{x}_0) = f(\mathbf{x}_0) \tag{7}$$

However $\mathbf{x}' = \theta\mathbf{x}_* + (1-\theta)\mathbf{x}_0$ is in the δ-neighbourhood of $\mathbf{x}$. This relation thus contradicts definition of local minimum. Therefore, $\forall \mathbf{x} \in \Omega$, $f(\mathbf{x}) \geq f(\mathbf{x}_0)$. In other words, the local minimum is also a global minimum. $\square$

(3) **Assuming sufficient smoothness, *strictly convex* functions must have positive semi-definite Hessians, but not necessarily positive definite.**

Counterinstance: $f(\mathbf{x}) = x_1^4 + x_2^4$. By definition f is indeed strictly convex, but $H_f(0) = \mathbf{0}$ is trivial, thus does not count as positive-definite. $\square$

Every *strictly convex* function has a unique global minimum.

Proof:

Let f be a *convex* function. Let $\mathbf{x}^*$ be its global minimum. Suppose $\exists \mathbf{x}' \neq \mathbf{x}^*$ so that $f(\mathbf{x}') = f(\mathbf{x}^*) = \min_{\mathbf{x} \in \Omega} f(\mathbf{x})$. By definition we have

$$f(\theta\mathbf{x}' + (1-\theta)\mathbf{x}^*) < \theta f(\mathbf{x}') + (1-\theta) f(\mathbf{x}^*) = \theta \min_{\mathbf{x} \in \Omega} f(\mathbf{x}) + (1-\theta) \min_{\mathbf{x} \in \Omega} f(\mathbf{x}) = \min_{\mathbf{x} \in \Omega} f(\mathbf{x}) \tag{8}$$

which contradicts the definition of global minimum. Therefore $\forall \mathbf{x} \neq \mathbf{x}^* \in \Omega$, $f(\mathbf{x}) > f(\mathbf{x}^*)$. In other words, the global minimum is unique. $\square$

(b) Properties of Hessian.

(1) **For non-convex function, its Hessian can be positive-(semi-)definite, negative-(semi-)definite or indefinite at different points.** There is no constraint that it be negative semi-definite.

Simplest counterinstance: $f(x) = x^3$. For $x \geq 0$ the Hessian is positive semi-definite, while for $x \leq 0$ it is negative semi-definite. $\square$

(2) **Positive-definiteness of the Hessian alone cannot determine optimality for convex functions.**

Simplest counterinstance: $f(x) = x^2$. The Hessian is positive-definite everywhere, but $x^* = 0$ is the only local and global minimum. $\square$

(3) **The function f has a local minimum at point $\mathbf{x}_0$ if $\nabla f(\mathbf{x}_0) = \mathbf{0}$ and the Hessian matrix is positive definite.** Positive determinant has not guaranteed implication as to the local geometry:

Simplest counterinstance: $f(\mathbf{x}) = -x_1^2 - x_2^2$. At $\mathbf{x} = 0$ we have:

$$\nabla f(\mathbf{0}) = \begin{pmatrix} -2x_1 \\ -2x_2 \end{pmatrix}_{\mathbf{0}} = \mathbf{0}, \quad H_f(\mathbf{x}) \equiv H_f(\mathbf{0}) = \begin{pmatrix} -2, -2 \end{pmatrix}, \quad \det\left(H_f\right) \equiv 4 > 0 \tag{9}$$

But apparently the function is concave and has no local minimum.

(c) Micellaneous

(1) **The set of all orthogonal $n \times n$ matrices is NOT a convex set in $\mathbb{R}^{n \times n}$.**

This is quite apparent, as the sum of different orthogonal basis does not yield another orthogonal basis.

(2) **$f(x_1, x_2) = \frac{1}{x_1 x_2}$ on $\mathbb{R}^2_{++}$ (all non-negative real numbers) is convex.**

Proof:

Consider any point inside the domain, we have:

$$H_f(x_1, x_2) = \begin{pmatrix} \frac{2}{x_1^3 x_2} & \frac{1}{x_1^2 x_2^2} \\ \frac{1}{x_1^2 x_2^2} & \frac{2}{x_1 x_2^3} \end{pmatrix} = \frac{1}{x_1^3 x_2^3} \begin{pmatrix} 2x_2^2 & x_1 x_2 \\ x_1 x_2 & 2x_1^2 \end{pmatrix} = \frac{1}{x_1^3 x_2^3} \left[\begin{pmatrix} x_2^2 & 0 \\ 0 & x_1^2 \end{pmatrix} + \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}^T \right] \tag{10}$$

The Hessian is positive definite every point within the domain, hence the function is convex. □

(3) **For the regularized linear least squares $f(\mathbf{w}) = \|\Phi\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2$ where $\lambda > 0$ and $\Phi \in \mathbb{R}^{n \times d}$, if $n \geq d$ and the columns of Φ are independent, then f has a unique global minimum.** However, with regularization, the condition for unique global minimum can be relaxed such that $\Phi^T\Phi + \lambda\mathbf{I} \succ 0$.

3 Linear regression

Consider the least squares misfit in two dimensional model space:

$$L(w_0, w_1) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2 \tag{11}$$

(a) Fixing $w_0 = 0$, the optimality condition for w_1 is computed by:

$$\frac{\partial L}{\partial w_1} = -\frac{1}{n} \sum_{i=1}^{n} x_i (y_i - w_1 x_i) = \frac{1}{n} \left[\left(\sum_{i=1}^{n} x_i^2 \right) w_1 - \sum_{i=1}^{n} x_i y_i \right] = 0 \tag{12}$$

This yields the following linear fit:

$$w_1^* = \arg\min_{w_1} L(0, w_1) = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \left(\sum_{i=1}^{n} x_i^2 \right)^{-1} \sum_{i=1}^{n} x_i y_i \tag{13}$$

(b) For $n \geq 2$ and $x_i \neq x_j$, $\forall 1 \leq i \neq j \leq n$, the Hessian is given by:

$$H_L(w_0, w_1) = \begin{pmatrix} \frac{\partial^2 L}{\partial w_0^2} & \frac{\partial^2 L}{\partial w_0 \partial w_1} \\ \frac{\partial^2 L}{\partial w_0 \partial w_1} & \frac{\partial^2 L}{\partial w_1^2} \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{n} \sum_{i=1}^n x_i \\ \frac{1}{n} \sum_{i=1}^n x_i & \frac{1}{n} \sum_{i=1}^n x_i^2 \end{pmatrix} \tag{14}$$

Since $\overline{x^2} - \bar{x}^2 = \frac{1}{n} \sum x_i^2 - (\frac{1}{n} \sum x_i)^2 > 0$ (using $x_i \neq x_j, i \neq j$), we have $\det(H_L) > 0$. And since the diagonal elements are positive, one can assert that $H_L \succ 0$. Therefore, the least squares function is convex with respect to $\mathbf{w} = (w_0, w_1)$. □

(c) For a strictly convex function, first order optimality condition suffices to guarantee the global minimum. The gradient:

$$\nabla L(w_0, w_1) = \begin{pmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \end{pmatrix} = \begin{pmatrix} -\frac{1}{n} \sum_{i=1}^n (y_i - w_0 - w_1 x_i) \\ -\frac{1}{n} \sum_{i=1}^n x_i (y_i - w_0 - w_1 x_i) \end{pmatrix} = \begin{pmatrix} w_0 + \overline{x} w_1 - \overline{y} \\ \overline{x} w_0 + \overline{x^2} w_1 - \overline{xy} \end{pmatrix} \tag{15}$$

(d) The optimal parameters are then given by solving:

$$\begin{aligned} w_0 + \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) w_1 &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) w_0 + \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) w_1 &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i \end{aligned} \tag{16}$$

The solution yields:

$$\begin{cases} w_0 = \dfrac{\left(\frac{1}{n}\sum x_i^2\right)\left(\frac{1}{n}\sum y_i\right) - \left(\frac{1}{n}\sum x_i\right)\left(\frac{1}{n}\sum x_i y_i\right)}{\left(\frac{1}{n}\sum x_i^2\right) - \left(\frac{1}{n}\sum x_i\right)^2} = \dfrac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} \\ w_1 = \dfrac{\frac{1}{n}\sum x_i y_i - \left(\frac{1}{n}\sum x_i\right)\left(\frac{1}{n}\sum y_i\right)}{\left(\frac{1}{n}\sum x_i^2\right) - \left(\frac{1}{n}\sum x_i\right)^2} = \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \end{cases} \tag{17}$$

This formulation can be easily extended to multi-dimensional variable space by converting the coefficients to matrices, where a linear least squares with intercept is formulated as follows:

$$L = \frac{1}{2n}\left\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\right\|^2, \qquad \mathbf{\Phi} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix} \tag{18}$$

where x_{ij} denotes the j-th variable of the i-th data point.

(e) $\mathbf{\Phi}^T\mathbf{\Phi}$ is invertible, if and only if $\mathbf{\Phi}$ is full column rank, i.e. $\operatorname{rank}(\mathbf{\Phi}) = d + 1$.

(f) Given $\mathbf{\Phi}^T\mathbf{\Phi}$ is invertible, one can immediately conclude that the Hessian is positive definite, and the objective function (loss function) is strictly convex. In this case the solution to the first-order optimality condition yields the global minimum:

$$\nabla L = \frac{1}{n}\left(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^T\mathbf{y}\right) = 0 \quad \Longrightarrow \quad \mathbf{w} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{19}$$

(g) If $n < d + 1$, then $\operatorname{rank}(\mathbf{\Phi}) \le n < d + 1$ and $\mathbf{\Phi}^T\mathbf{\Phi}$ and must contain zero eigenvalue. Consider the non-trivial space formed by eigenvectors of $\mathbf{\Phi}^T\mathbf{\Phi}$ corresponding to trivial eigenvalues, equivalently the nullspace of $\mathbf{\Phi}^T\mathbf{\Phi}$. Assuming $\mathbf{w}_0$ satisfies the first-order optimality condition, we have

$$\mathbf{\Phi}^T\mathbf{\Phi}(\mathbf{w}_1) = \mathbf{\Phi}^T\mathbf{\Phi}(\mathbf{w}_0 + \mathbf{w}') = \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}_0 + \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}' = \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}_0 + \mathbf{0} = \mathbf{\Phi}^T\mathbf{y}, \qquad \forall \mathbf{w}' \in \ker(\mathbf{\Phi}^T\mathbf{\Phi}) \tag{20}$$

so $\mathbf{w}_1 = \mathbf{w}_0 + \mathbf{w}'$ also satisfies the optimality condition. However, as $\mathbf{\Phi}^T\mathbf{\Phi} \succeq 0$, the function is still convex, and any point fulfilling the first-order optimality is a global minimum. Therefore, there are infinitely many global minima for function L.

Consider using the gradient descent scheme on linear least squares:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\nabla L(\mathbf{w}), \qquad \left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|_2 \le \left\|\mathbf{I} - \eta\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{op}\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 \tag{21}$$

(h) Assuming the stepsize η is such that $\left\|\mathbf{I} - \eta\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{op} < 1$. A total number of

$$\tau = \left\lceil \log_{\left\|\mathbf{I} - \eta\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{op}} \frac{\varepsilon}{\left\|\mathbf{w}^0 - \mathbf{w}^*\right\|_2} \right\rceil = \left\lceil \frac{\ln\varepsilon - \ln\left\|\mathbf{w}^0 - \mathbf{w}^*\right\|_2}{\ln\left\|\mathbf{I} - \eta\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{op}} \right\rceil \tag{22}$$

is needed to guarantee that the solution is within the ε-neighbourhood of the ground truth.

(i) Let λ_{max} and λ_{min} be the maximum and minimum eigenvalues of $\mathbf{\Phi}^T\mathbf{\Phi}$. Then the spectrum of $\mathbf{I} - \eta\mathbf{\Phi}^T\Phi$ is bounded by $1 - \eta\lambda_{max}$ and $1 - \eta\lambda_{min}$, and $\left\|\mathbf{I} - \eta\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{op} = \max\{|1 - \eta\lambda_{max}|, |1 - \eta\lambda_{min}|\}$.

To minimize this convergence rate, one can consider η as the independent variable and the convergence rate $p = \left\|\mathbf{I} - \eta\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{op}$ as a function of it. Assuming $\lambda_{max} \ge \lambda_{min} > 0$ so that a unique $\mathbf{w}^*$ exists, the optimal stepsize can be determined via $1 - \eta\lambda_{min} = \eta\lambda_{max} - 1$:

$$\eta^* = \frac{2}{\lambda_{max} + \lambda_{min}}, \quad p^* = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{(\lambda_{max}/\lambda_{min}) - 1}{(\lambda_{max}/\lambda_{min}) + 1} = \frac{\operatorname{cond}(\mathbf{\Phi}^T\mathbf{\Phi}) - 1}{\operatorname{cond}(\mathbf{\Phi}^T\mathbf{\Phi}) + 1} \tag{23}$$

In particular, when $\lambda_{max} = \lambda_{min} = \lambda$, in other words the loss function is basically isotropic, one will arrive at:

$$\eta^* = \frac{1}{\lambda}, \qquad p^* = 0 \tag{24}$$

meaning that with this optimal step size, the algorithm reaches the global minimum in one iteration. This is nothing surprising, for the negative gradients of such isotropic loss function always point exactly towards the global minimum.

(j) In closed form, the least squares can be solved by solving the normal equation once. Evaluating the matrix and vector requires $\approx nd^2$ multiplications and additions. In the case of full-column-rank $\mathbf{\Phi}$ and so strictly convex loss function with moderate dimensions of variables, the system can be best solved by direct solvers such as Cholesky factorization. In all, this requires around $\frac{1}{6}d^3 + nd^2$ multiplications and additions.

Using steepest descent method, however, at each iteration the evaluation of the gradient requires around $3nd$ multiplications and additions, and the total expense would be around $3\tau nd$. Use the explicit expression for τ, the expense can be rewritten as:

$$3\tau nd \approx \frac{\ln \frac{\left\|\mathbf{w}^0 - \mathbf{w}^*\right\|}{\varepsilon}}{\ln \frac{\mathrm{cond}(\Phi^{\mathrm{T}}\Phi)+1}{cond(\mathbf{\Phi}^T\mathbf{\Phi})-1}} 3nd \tag{25}$$

Hence the complexity depends on the convergence threshold, the initial estimate, and the conditioning of the Hessian. In particular, denoting $s = \ln[\left\|\mathbf{w}^0 - \mathbf{w}^*\right\| / \varepsilon]$ as some parameter determined by initial guess and convergence threshold, when we consider relatively large condition numbers, it can be approximated as:

$$3\tau nd \approx \frac{3}{2}\mathrm{cond}(\mathbf{\Phi}^T\mathbf{\Phi})snd \tag{26}$$

Since $n > d$, the complexity for direct solving normal equation is mostly dominated by nd^2. Therefore, when $\mathrm{cond}(\mathbf{\Phi}^T\mathbf{\Phi}) > d/s$, gradient descent may be more expensive than solving normal equation; otherwise gradient descent may be more favourable.

(k) See the code below.

(l) See the code below.

```python
import numpy as np

# Data points
xy = [[-3.4, 7.5], [-0.9, 2.3], [2.9, 8.7], [4, -6.9], [4.9, -18.9]]
xy = np.asarray(xy)

# Construct basis and matrix
fit_func = [lambda x: np.ones(x.shape),
            lambda x: x,
            lambda x: x**2,
            np.sin]
Phi = np.ones((xy.shape[0], len(fit_func)))
for i in range(len(fit_func)):
    Phi[:, i] = fit_func[i](xy[:, 0])

# Solve normal eqn.
coeffs = np.linalg.solve(Phi.T @ Phi, Phi.T @ xy[:, 1])
print(coeffs)                   # coefficients
print(Phi @ coeffs)             # recovery
print(xy[:, 1] - Phi @ coeffs)  # residual
```

4 Gradient descent

The loss function is defined by:

$$L(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{X}\mathbf{w} - \mathbf{y}\right\|_2^2 \tag{27}$$

The gradient descent scheme:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \eta\nabla L(\mathbf{w}^k) \tag{28}$$

(a) Plug in the expression for L,

$$\nabla L(\mathbf{w}^k) = \mathbf{X}^T\mathbf{X}\mathbf{w}^k - \mathbf{X}^T\mathbf{y}, \qquad \mathbf{w}^{k+1} = \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)\mathbf{w}^k + \eta\mathbf{X}^T\mathbf{y} \tag{29}$$

(b) Assuming at certain k, $\mathbf{w}^k$ satisfies:

$$\mathbf{w}^k = \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^k \mathbf{w}^0 + \eta\left(\sum_{j=0}^{k-1}\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^j\right)\mathbf{X}^T\mathbf{y} \tag{30}$$

Then at the $()k+1$)-th iteration, the updated estimate can be written as:

$$\begin{aligned}
\mathbf{w}^{k+1} &= \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)\mathbf{w}^k + \eta\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)\left[\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^k\mathbf{w}^0 + \eta\sum_{j=0}^{k-1}\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^j\mathbf{X}^T\mathbf{y}\right] + \eta\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^{k+1}\mathbf{w}^0 + \eta\sum_{j=1}^{k}\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^j\mathbf{X}^T\mathbf{y} + \eta\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^{k+1}\mathbf{w}^0 + \eta\left[1 + \sum_{j=1}^{k}\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^j\right]\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^{k+1}\mathbf{w}^0 + \eta\sum_{j=0}^{k}\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^j\mathbf{X}^T\mathbf{y}
\end{aligned} \tag{31}$$

also satisfying the same relation. Now that $\mathbf{w}^1 = \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^1\mathbf{w}^0 + \eta\sum_{j=0}^{0}\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^j\mathbf{X}^T\mathbf{y}$ satisfies the relation, using mathematical induction, one can conclude:

$$\mathbf{w}^k = \left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^k\mathbf{w}^0 + \eta\sum_{j=0}^{k-1}\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^j\mathbf{X}^T\mathbf{y}, \qquad \forall k \in \mathbb{Z}^+ \tag{32}$$

(c) The eigenvalues of $\mathbf{X}^T\mathbf{X}$ can be constructed via the SVD of $\mathbf{X}$:

$$\mathbf{X}^T\mathbf{X} = \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\right)^T\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\right) = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{33}$$

where $\Lambda_{ij} = \sigma_i^2\delta_{ij}$. Therefore:

$$\mathbf{I} - \eta\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{V}^T - \eta\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T = \mathbf{V}\left(\mathbf{I} - \eta\mathbf{\Lambda}\right)\mathbf{V}^T \tag{34}$$

(d) No question.

(e) Using the eigenvalue formulation, the power of a matrix can be easily computed:

$$\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^k = \left(\mathbf{V}(\mathbf{I} - \eta\mathbf{\Lambda})\mathbf{V}^T\right)^k = \mathbf{V}(\mathbf{I} - \eta\mathbf{\Lambda})^k\mathbf{V}^T \tag{35}$$

(f) If $\mathbf{v}^i$ is an eigenvector of $\mathbf{X}^T\mathbf{X}$ corresponding to eigenvalue σ_i^2, we have

$$\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^k\mathbf{v}^i = \mathbf{V}\left(\mathbf{I} - \eta\mathbf{\Lambda}\right)^k\mathbf{V}^T\mathbf{v}^i = \mathbf{V}\left(\mathbf{I} - \eta\mathbf{\Lambda}\right)^k\mathbf{e}_i = (1 - \eta\sigma_i^2)^k\mathbf{v}^i \tag{36}$$

If $\sigma_i > 0$, $\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^k\mathbf{v}^i = (1 - \eta\sigma_i^2)^k\mathbf{v}^i \to \mathbf{0}$ $(k \to +\infty)$;

If $\sigma_i = 0$, $\left(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X}\right)^k\mathbf{v}^i = (1 - \eta\sigma_i^2)^k\mathbf{v}^i \equiv \mathbf{v}^i$.

(g) Therefore the first term virtually extracts the part of $\mathbf{w}^0$ that is inside the nullspace of $\mathbf{X}^T$. It can also be interpreted as $\mathbf{w}^0$ being squeezed onto the orthogonal complement of $\mathbf{X}^+$ each iteration.

(h) Similarly, utilizing the eigenvalue decomposition:

$$\eta\sum_{j=0}^{k-1}(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X})^j\mathbf{X}^T\mathbf{y} = \eta\mathbf{V}\sum_{j=0}^{k-1}\left(\mathbf{I} - \eta\mathbf{\Lambda}\right)^j\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{y} \tag{37}$$

(i) First we compute the middle part, i.e. the matrix series. Diagonality is preserved during matrix multiplications and summations, and the i-th diagonal component of the resulting matrix:

$$\begin{aligned}\left(\sum_{j=0}^{k=1}(\mathbf{I}-\eta\mathbf{\Lambda})^j\right)_{ii} &= \sum_{j=0}^{k-1}(1-\eta\sigma_i^2)^j \to \frac{1}{\eta\sigma_i^2} \qquad k\to\infty \qquad (\sigma_i\neq 0)\\ \left(\sum_{j=0}^{k=1}(\mathbf{I}-\eta\mathbf{\Lambda})^j\,\mathbf{\Sigma}^T\right)_{ii} &= \sigma_i\sum_{j=0}^{k-1}(1-\eta\sigma_i^2)^j \to \frac{1}{\eta\sigma_i} \qquad k\to\infty \qquad (\sigma_i\neq 0)\end{aligned} \tag{38}$$

In principle, those diagonal elements corresponding to nullspace will diverge when $k\to\infty$. However, as they are multiplied by $\mathbf{\Sigma}^T$, whose respective diagonal elements are zero, they will effectively not contribute to the entire product:

$$\left(\sum_{j=0}^{k=1}(\mathbf{I}-\eta\mathbf{\Lambda})^j\,\mathbf{\Sigma}^T\right)_{ii} = \sigma_i\sum_{j=0}^{k-1}(1-\eta\sigma_i^2)^j \equiv 0 \qquad (\sigma_i=0) \tag{39}$$

Denote $\mathbf{\Sigma}^+\in\mathbb{R}^{d\times n}$ as the inverse of $\mathbf{\Sigma}$ with truncated singular values, i.e.

$$\mathbf{\Sigma}^+=\sigma_i^+\delta_{ij}, \qquad \sigma_i^+=\begin{cases}\sigma_i^{-1} & (\sigma_i\neq 0)\\ 0 & (\sigma_i=0)\end{cases} \tag{40}$$

The 2nd term can be reiterated as:

$$\eta\mathbf{V}\sum_{j=0}^{k-1}(\mathbf{I}-\eta\mathbf{\Lambda})^j\,\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{y} = \eta\mathbf{V}\frac{1}{\eta}\mathbf{\Sigma}^+\mathbf{U}^T\mathbf{y} = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^T\mathbf{y} \tag{41}$$

Similarly one can see if we split $\mathbf{y}$ into $\ker(\mathbf{X}^T)$ and $\ker(\mathbf{X}^T)^\perp$ parts, those inside the nullspace will get squeezed to zero, and those outside the nullspace will be scaled by inverse of the singular value.

(j) We see that $\mathbf{X}^-=\mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^T$ has the following properties:

$$\begin{aligned}&\mathbf{X}^-\mathbf{X}\mathbf{X}^- = \mathbf{V}\mathbf{\Sigma}^+\mathbf{\Sigma}\mathbf{\Sigma}^+\mathbf{U}^T = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^T = \mathbf{X}^-\\ &\mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^+\mathbf{\Sigma}\mathbf{V}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{X}\\ &\left(\mathbf{X}\mathbf{X}^-\right)^T = \left(\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^+\mathbf{U}^T\right)^T = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^+\mathbf{U}^T = \mathbf{X}\mathbf{X}^-\\ &\left(\mathbf{X}^-\mathbf{X}\right)^T = \left(\mathbf{V}\mathbf{\Sigma}^+\mathbf{\Sigma}\mathbf{V}^T\right)^T = \mathbf{V}\mathbf{\Sigma}^+\mathbf{\Sigma}\mathbf{V}^T = \mathbf{X}^-\mathbf{X}\end{aligned} \tag{42}$$

Therefore $\mathbf{X}^+=\mathbf{X}^-=\mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^T$ is the Moore-Penrose inverse of $\mathbf{X}$, and the estimate at the k-th iteration:

$$\begin{aligned}\mathbf{w}^k &\to \mathbf{V}\begin{pmatrix}\mathbf{0}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{d-r}\end{pmatrix}\mathbf{V}^T\mathbf{w}^0 + \mathbf{X}^+\mathbf{y}\\ &= \left(\mathbf{I}-\mathbf{X}^+\mathbf{X}\right)\mathbf{w}^0 + \mathbf{X}^+\mathbf{y} = \mathbf{w}^0 + \mathbf{X}^+\left(\mathbf{y}-\mathbf{X}\mathbf{w}^0\right) \qquad (k\to+\infty)\end{aligned} \tag{43}$$

where r denotes the rank of $\mathbf{X}^T\mathbf{X}$.

(k) For under-parameterized (equiv. over-determined) setting, $r=d$. In this case $\dim\ker(\mathbf{X})=0$, and $\mathbf{w}^k\to\mathbf{X}^+\mathbf{y}$ $(k\to+\infty)$ regardless of choice of $\mathbf{w}^0$.