MML lecture extra notes, week Oct 25 - 29, 2021

Linear regression considers solving the following task:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}), \quad L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2. \tag{1}$$

Arithmetico–geometric sequence

In linear regression, gradient descent returns an update rule as

$$\boldsymbol{\theta}_{t+1} = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})\boldsymbol{\theta}_t + \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y}. \tag{2}$$

The solution of this iterative update is related to an arithmetico–geometric sequence. Writing $\boldsymbol{\theta}_{t+1} + \boldsymbol{\beta} = \mathbf{A}(\boldsymbol{\theta}_t + \boldsymbol{\beta})$ with $\mathbf{A} := \mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X}$, we would like to solve for $\boldsymbol{\beta}$ such that:

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} = \mathbf{A}(\boldsymbol{\theta}_t + \boldsymbol{\beta}) - \boldsymbol{\beta} &= (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})\boldsymbol{\theta}_t + \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y} \\
\Leftrightarrow \quad -\frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y} \\
\Leftrightarrow \quad \boldsymbol{\beta} &= -(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = -\boldsymbol{\theta}^*.
\end{aligned} \tag{3}$$

So this immediately implies

$$\boldsymbol{\theta}_t - \boldsymbol{\theta}^* = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) \quad \Rightarrow \quad \boldsymbol{\theta}_t = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + \boldsymbol{\theta}^*, \tag{4}$$

which means $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ if $(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) \to \mathbf{0}$.

Rayleigh quotient

Assume $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$ is symmetric (so that also $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^\top$). Consider the following *Rayleigh quotient*

$$R(\mathbf{A}, \boldsymbol{x}) = \frac{\boldsymbol{x}^\top\mathbf{A}\boldsymbol{x}}{||\boldsymbol{x}||_2^2}, \quad ||\boldsymbol{x}||_2^2 = \boldsymbol{x}^\top\boldsymbol{x}. \tag{5}$$

Using the fact that $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$, we can define $\boldsymbol{z} = \mathbf{Q}^\top\boldsymbol{x}$ and rewrite the Rayleigh quotient as:

$$R(\mathbf{A}, \boldsymbol{x}) = \frac{\boldsymbol{x}^\top\mathbf{Q}\Lambda\mathbf{Q}^\top\boldsymbol{x}}{\boldsymbol{x}^\top\mathbf{Q}\mathbf{Q}^\top\boldsymbol{x}} = \frac{\boldsymbol{z}^\top\Lambda\boldsymbol{z}}{\boldsymbol{z}^\top\boldsymbol{z}}. \tag{6}$$

As $\Lambda = \text{diag}(\lambda_1, ..., \lambda_D)$ is a diagonal matrix, we have (writing $\boldsymbol{z} = (z_1, ..., z_D)^\top$)

$$\boldsymbol{z}^\top\Lambda\boldsymbol{z} = \sum_{d=1}^{D}\lambda_d z_d^2. \tag{7}$$

Therefore the Rayleigh quotient can be written as the following weighted average of the eigenvalues

$$R(\mathbf{A}, \boldsymbol{x}) = \sum_{d=1}^{D}\frac{z_d^2}{||\boldsymbol{z}||_2^2}\lambda_d, \quad \text{with} \quad \sum_{d=1}^{D}\frac{z_d^2}{||\boldsymbol{z}||_2^2} = 1. \tag{8}$$

In summary, these derivation indicate that the Rayleigh quotient is bounded as

$$\begin{aligned}
&\lambda_{min}(\mathbf{A}) \leq R(\mathbf{A}, \boldsymbol{x}) \leq \lambda_{max}(\mathbf{A}) \\
\Rightarrow \quad &\lambda_{min}(\mathbf{A})||\boldsymbol{x}||_2^2 \leq \boldsymbol{x}^\top\mathbf{A}\boldsymbol{x} \leq \lambda_{max}(\mathbf{A})||\boldsymbol{x}||_2^2,
\end{aligned} \tag{9}$$

where $\lambda_{min}(\mathbf{A})$ and $\lambda_{max}(\mathbf{A})$ are the smallest and largest eigenvalues of $\mathbf{A}$, respectively.

An extra exercise

Show that solving linear regression using gradient descent with momentum, if converges, converges to $\boldsymbol{\theta}^* = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$.

Hint: consider the simpler case with fixed step size γ and momentum factor α. Follow the below steps and practice your linear algebra skills :)

1. Write down the update equations for the parameters $\boldsymbol{\theta}_t$ and the momentum $\Delta\boldsymbol{\theta}_t$;
2. Collect both terms as a long vector $(\boldsymbol{\theta}_t^\top, \Delta\boldsymbol{\theta}_t^\top)^\top$, and merge the two linear update equations in step 1 into one "joint" linear equation using block matrices;
3. Apply the analysis techniques in GD for linear regression to show the converged solution (if converges).

(The solution will be uploaded shortly.)