Lecture Notes IV.2 – Simple analysis of gradient descent

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

October, 2021

Rate of linear convergence

Newton-Raphson "rounds" the surface of f around minimum

Implicit bias of Gradient Descent

Reading HTF Ch.: –, Murphy Ch.: –, Bach Chapter 5.2, 10.1

Useful facts

Assume that our function f is quadratic, i.e

$$f(x) = \frac{1}{2}x^T Hx + g^T x + c \ \text{ with } H \succ 0. \tag{1}$$

Then,

$$\begin{aligned} \nabla f(x) &= Hx + g = H(x - x^*) && (2) \\ \nabla^2 f(x) &= H && (3) \\ x^* &= -H^{-1}g, \quad \text{and} \quad Hx^* = -g && (4) \\ & && (5) \end{aligned}$$

Gradient descent $x^{t+1} = x^t - \eta \nabla f(x^t)$

Rate of linear convergence

$$\begin{aligned}
x^{t+1} - x^* &= (x^t - \eta H(x^t - x^*)) - x^* && (6)\\
&= [I - \eta H](x^t - x^*) \;=\; (I - \eta H)^t (x^0 - x^*) && (7)\\
e^{t+1} &\leq \|I - \eta H\|^t e^0 \quad \text{with } e^t = \|x^t - x^*\| && (8)\\
f(x) - f(x^*) &= \frac{1}{2}(x - x^*)^T H (x - x^*) \quad \text{for any } x && (9)\\
\text{Proof} & && \\
\frac{1}{2}(x - x^*)^T H (x - x^*) &= \frac{1}{2}x^T H x + \frac{1}{2}(x^*)^T H x^* - \underbrace{x^T H x^*}_{-x^T g} \quad \text{recall } Hx^* = -g && (10)\\
&= f(x) - \left(\frac{1}{2}(x^*)^T H x^* + g^T x^*\right) && (11)\\
\text{Hence,} & && \\
f(x) - f(x^*) &= \frac{1}{2}(x^0 - x^*)^T (I - \eta H)^{2t} H (x^0 - x^*) && (12)\\
\text{because} & \quad H(I - \eta H) = (I - \eta H)H && (13)
\end{aligned}$$

Choice of η

For convergence, we want to control the maximum eigenvalue of $(I - \eta H)$. Let m, M the min, max singular values of H.

$$\text{minimize}_{\eta} \; max_{\lambda \in [m,M]} |1 - \eta\lambda| \tag{14}$$

We obtain $\frac{1}{\eta^*} = \frac{M+m}{2}$ or

$$\eta^* = \frac{2}{M+m} \tag{15}$$

For this η^* we obtain

$$\beta^* \equiv \sigma_{max}(I - \eta H) \;=\; \frac{M-m}{M+m} \tag{16}$$

This value is always in $[0,1]$. Denote by $\kappa = \frac{M}{m}$ the **condition number** of H; β^* approaches 1 when κ is large.

Newton-Raphson "rounds" the surface of f around minimum

- If we take $H = I$, then $\beta = 0$, meaning that the first order convergence is infinitely fast (super-linear convergence).
- How can we make $H = I$? We transform the variable x by

$$x \;=\; H^{-1/2}z, \quad z \;=\; H^{1/2}x \tag{17}$$

Then $f(z) = \frac{1}{2}\|z\|^2 + g^T H^{-1/2}z + c$ and the new Hessian is I.
Let us look at the gradient descent in z.

$$\begin{aligned}
\nabla_z f(z) &= z + (H^{-1/2})^T g & (18)\\
z^{t+1} &= z^t - \eta(z^t + (H^{-1/2})^T g) & (19)\\
x^{t+1} &= H^{-1/2}z^{t+1} = (1-\eta)H^{-1/2}z^t - \eta H^{-1}g & (20)\\
&= (1-\eta)x^t - \eta \underbrace{\nabla^2_x f(x^t)\nabla_x f(x^t)}_{\text{Newton step}} & (21)
\end{aligned}$$

- Hence the Newton step is a gradient step in the transformed coordinates z.

For a symmetric $A \succ 0$, $B = A^{1/2}$ is a matrix for which $B^T B = A$ holds; $A^{1/2}$ is not unique. We have also $A^{-1} = (B^T B)^{-1} = B^{-1}(B^T)^{-1}$. Exercise Prove that B is non-singular when A is non-singular; find the equivalence class of all B which are the square root of some A.

Gradient descent for Least Squares Loss

Consider linear regression, with $f(\theta) \equiv L_{LS}(\theta) = \frac{1}{2n}\|y - \mathbf{X}\theta\|^2$ with $d < n$. Let $\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}$ be the **kernel matrix** and $H = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ the **covariance** matrix.

$$f(\theta) \;=\; \frac{1}{2}\theta^T H\theta - \underbrace{\frac{1}{n}y^T\mathbf{X}}_{g}\,\theta + \frac{1}{2n}y^T y \tag{22}$$

- We start from $\theta^0 = 0$.
- We don't assume the solution is unique. In other words, H may be singular.
- In particular, note that for $d > n$, H is singular, but K is invertible w.l.o.g. when the system $\mathbf{X}\theta = y$ has a solution (and the system has an infinite number of solutions).
- For any θ^* satisfying $y = \mathbf{X}\theta^*$ and for some iterate θ^t we have

$$\theta^t - \theta^* \;=\; (I - \eta H)^t(\theta^0 - \theta^*) \tag{23}$$

$$\theta^t \;=\; [I - (I - \eta H)^t]\theta^* \tag{24}$$

The GD path

- Now on the GD path (which is deterministic given $\mathbf{X}$)

$$\nabla f(0) = g = \frac{1}{n}\mathbf{X}^T y \tag{25}$$

$$\theta^1 = 0 - \eta \nabla f(0) = -\eta \frac{1}{n}\mathbf{X}^T y \tag{26}$$

  Thus θ^1 is a linear combination of the rows of $\mathbf{X}$ (i.e. of the data points).
- By induction, θ^t for any t is a linear combination of the rows of $\mathbf{X}$, hence

$$\theta^t = \mathbf{X}^T \alpha^t, \quad \text{with } \alpha^t \in \mathbb{R}^n \tag{27}$$

- Since the gradient is non-zero whenever $y \neq \mathbf{X}\theta$, the GD algorithm converges to a point[1] where $y = \mathbf{X}\theta = \mathbf{X}\mathbf{X}^T \alpha$.
- When K is invertible, let $\alpha^* = K^{-1} y$; then $\theta^* = \mathbf{X}^T \alpha^*$ is the limit of GD.

[1] This is informal. What we can say that when t is sufficiently large, $\mathbf{X}\theta^t = \mathbf{X}\mathbf{X}^T\alpha^t$ is arbitrarily close to y.

θ^* is the minimum norm solution of $\mathbf{X}\theta = y$

- To prove this, we must use convex duality.

$$\text{Primal: } \inf_{\theta} \frac{1}{2}\|\theta\|^2 \text{ s.t. } \mathbf{X}\theta = y \quad \Leftrightarrow \quad \text{Dual: } \sup_{\alpha} \inf_{\theta} \frac{1}{2}\|\theta\|^2 + \alpha^T(y - \mathbf{X}\theta) \tag{28}$$

- Solving the optimization over θ as a function of the parameter α we obtain $\theta = \mathbf{X}^T\alpha$.
- We replace θ in (28) to obtain

$$\sup_{\alpha} \alpha^T y - \frac{1}{2}\alpha^T K\alpha \tag{29}$$

  This is a concave function with optimum $\alpha^* = K^{-1}y$ **Yes, we get the same α^* from the previous page!**
- Finally, the solution to the Primal problem is $\theta^* = \mathbf{X}^T\alpha^* = \mathbf{X}^T K^{-1}y$, the solution obtained by Gradient Descent!

Note that θ^* above is not the OLS solution. In OLS, we minimize residuals norm, here we minimize the θ norm.