Solution 1b

BY XIAONAN PENG

Qishi Intermediate ML Study Group

Problem.

$$\hat{\beta} = \arg\min_b \left\{ \|b\|_2 \colon b \text{ minimizes } \frac{1}{2n}\|y - Xb\|_2^2 \right\}$$

a)

Case 1: $n \geqslant p = r(X)$, normal linear regression, we know the BLUE is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Case 2: $p > n = r(X)$, high-dimensional case. From information of rank, we know that the linear equation $Xb = y$ always has infinity solutions. Our goal is find the minimum solution. Since X has full row rank, it has right inverse

$$X^\dagger = X^T (XX^T)^{-1}$$

where XX^T is invertible since $r(XX^T) = r(X) = n$. It's easy to see that $X^\dagger y$ is a solution. We only need to prove it has the minimum norm. Let $\hat{\beta} = X^T(XX^T)^{-1}y$, β be a solution of $X\beta = y$.

$$\begin{aligned} (\beta - \hat{\beta})^T \hat{\beta} &= (\beta - \hat{\beta})^T X^T (XX^T)^{-1} y \\ &= (X(\beta - \hat{\beta}))^T (XX^T)^{-1} y \\ &= 0 \end{aligned}$$

Then we have $(\beta - \hat{\beta}) \perp \hat{\beta}$, $\hat{\beta}$ is the minimal norm solution.

We can compute the degrees of freedom:

$$n \geqslant p \colon \sum_i \frac{\partial \hat{y}_i}{\partial y_i} = \operatorname{tr}(X(X^T X)^{-1} X^T) = p$$

$$n < p \colon \sum_i \frac{\partial \hat{y}_i}{\partial y_i} = \operatorname{tr}(XX^T (XX^T)^{-1}) = n$$

b)

Gradient descent: $\beta^k = \beta^{k-1} - \varepsilon \nabla L(\beta^{k-1}) = \beta^{k-1} + \varepsilon \frac{X^T}{n}(y - X\beta^{k-1})$. Suppose $\hat{\beta}$ is the optimal solution which satisfies first order condition: $\nabla L(\hat{\beta}) = \frac{X^T}{n}(y - X\hat{\beta}) = 0$, then we have

$$\begin{aligned} \|\beta^k - \hat{\beta}\|_2 &= \left\| \beta^{k-1} + \varepsilon \frac{X^T}{n}(y - X\beta^{k-1}) - \hat{\beta} \right\|_2 \\ &= \left\| \left(I - \varepsilon \frac{X^T X}{n} \right)(\beta^{k-1} - \hat{\beta}) \right\|_2 \\ &\leqslant \left\| I - \varepsilon \frac{X^T X}{n} \right\|_2 \|\beta^{k-1} - \hat{\beta}\|_2 \\ &\leqslant \max\left(|1 - \varepsilon\lambda_{\min}|, |1 - \varepsilon\lambda_{\max}|\right) \|\beta^{k-1} - \hat{\beta}\|_2 \end{aligned}$$

where λ is the eigenvalue of $\frac{1}{n}X^T X$. We can see that the convergence factor is $\max\left(|1 - \varepsilon\lambda_{\min}|, |1 - \varepsilon\lambda_{\max}|\right)$, if GD converges to $\hat{\beta}$, then

$$\begin{aligned} \max\left(|1 - \varepsilon\lambda_{\min}|, |1 - \varepsilon\lambda_{\max}|\right) &< 1 \\ \Rightarrow 1 - \varepsilon\lambda &\in (-1, 1) \\ \Rightarrow \varepsilon\lambda &\in (0, 2) \end{aligned}$$

Since $\frac{1}{n}X^T X$ is semi-definite, $\lambda \geqslant 0$, then we just need $\varepsilon < \frac{2}{\lambda_{\max}}$ this one condition. ($\lambda_{\max}$ must be positive, otherwise $X^T X = 0$). Therefore, we can see that if $\varepsilon \in \left(0, \frac{1}{\lambda_{\max}}\right)$, GD must converge to $\hat{\beta}$, which satisfies normal equation: $X^T X\hat{\beta} = X^T y$.

If X has full column rank, then the solution is unique

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

is the exact solution in (a).

If X has full row rank, then we need to consider the initialization β^0.

$$\beta^k = \beta^{k-1} + \varepsilon \frac{X^T}{n}(y - X\beta^{k-1})$$

Notice that GD actually add a new vector in $C(X^T)$ in each step, so we can write $\beta^k = \beta^0 + X^T u^k$ for some vector sequence $\{u^k\}$

There are two cases:

- $\beta^0 \in C(X^T)$, then we can write $\beta^0 = X^T v$ and $\beta^k = X^T(v + u^k)$ is always in $C(X^T)$. Let $k \to \infty$, $\beta^k \to \hat{\beta}$ is also in $C(X^T)$. Then we have $\hat{\beta} = X^T(v + \hat{u})$. Then

  $$X\hat{\beta} - y = XX^T(v + \hat{u}) - y = 0 \Rightarrow v + \hat{u} = (XX^T)^{-1} y \Rightarrow \hat{\beta} = X^T(XX^T)^{-1} y$$

  which is the minimum norm solution in (a).

- $\beta^0 \notin C(X^T)$, we can write $\hat{\beta} = \beta^0 + X^T \hat{u}$. It's easy to see that $\hat{\beta} \notin C(X^T)$, then it can not be the minimal norm solution.

c)

We consider the usual least squares regression problem

$$\min_{\beta} \frac{1}{2n} \|y - X\beta\|_2^2 \tag{1}$$

where $X \in \mathbb{R}^{n \times p}$ is the design matrix, $y \in \mathbb{R}^n$ is the response. The gradient descent iteration is

$$\beta^k = \beta^{k-1} - \frac{\varepsilon}{n} X^T (X\beta - y) \tag{2}$$

where ε is a fixed learning rate. Let $\varepsilon \to dt$, we have the continuous form version: gradient flow.

$$d\beta(t) = \frac{1}{n} X^T (y - X\beta(t)) dt \tag{3}$$

to see the connection between (2) and (3), we simply rearrange (2) to be

$$\frac{\beta^k - \beta^{k-1}}{\varepsilon} = \frac{1}{n} X^T (y - X\beta^{k-1})$$

and setting $\beta(t) = \beta^k$, with $t = k\varepsilon$. The gradient flow equation is easy to solve under given initial condition $\beta(0) = 0$, which has the unique solution

$$\hat{\beta}^{\mathrm{gf}}(t) = (X^T X)^{\dagger} (I - \exp(-t X^T X / n)) X^T y \tag{4}$$

where $(X^T X)^{\dagger}$ is the Moore-Penrose inverse(Pseudo inverse) of $X^T X$.

The matrix expoential is defined as $\exp(A) = I + A + A^2/2 + A^3/3! + \cdots$

We can compute a simplier form, let $\frac{X^T X}{n} = U D U^T$, we have

$$\begin{aligned}
I - \exp(-t X^T X / n) &= I - U \exp(-tD) U^T \\
&= U(I - \exp(-tD)) U^T \\
\Rightarrow \ (X^T X)^{-1} &= \frac{1}{n} U D^{-1} U^T \\
\Rightarrow \ \hat{\beta}^{\mathrm{gf}}(t) &= \frac{1}{n} U D^{-1} U^T U (I - \exp(-tD)) U^T X^T y \\
&= \frac{1}{n} U \left(D^{-1} - D^{-1} \exp(-tD)\right) U^T X^T y
\end{aligned}$$

d)

$$\begin{aligned}\sum_i \frac{\partial \hat{y}_i}{\partial y_i} &= \mathrm{tr}(X(X^TX+\lambda I)^{-1}X^T) \\ &= \mathrm{tr}(UDV^T(VD^T\mathrm{D}\mathrm{V}^T+\lambda I)^{-1}VD^TU^T) \\ &= \mathrm{tr}(UD(\,D^TD+\lambda I)^{-1}D^TU^T) \\ &= \mathrm{tr}(D^TD(\,D^TD+\lambda I)^{-1}) \\ &= \sum_{i=1}^{p}\frac{d_i^2}{d_i^2+\lambda} < p\end{aligned}$$

e)

We want to compare the gradient flow and ridge regression. Let $X=\sqrt{n}US^{\frac{1}{2}}V^T$, so that $X^TX/n=VSV^T$, then we get

$$\begin{aligned}X\hat{\beta}^{\mathrm{ridge}}(\lambda) &= US(S+\lambda I)^{-1}U^Ty \\ X\hat{\beta}^{\mathrm{gf}}(t) &= U(I-\exp(-tS))U^Ty\end{aligned}$$

Let s_i be the i-th singular value of X, we have two shrinkage operator:

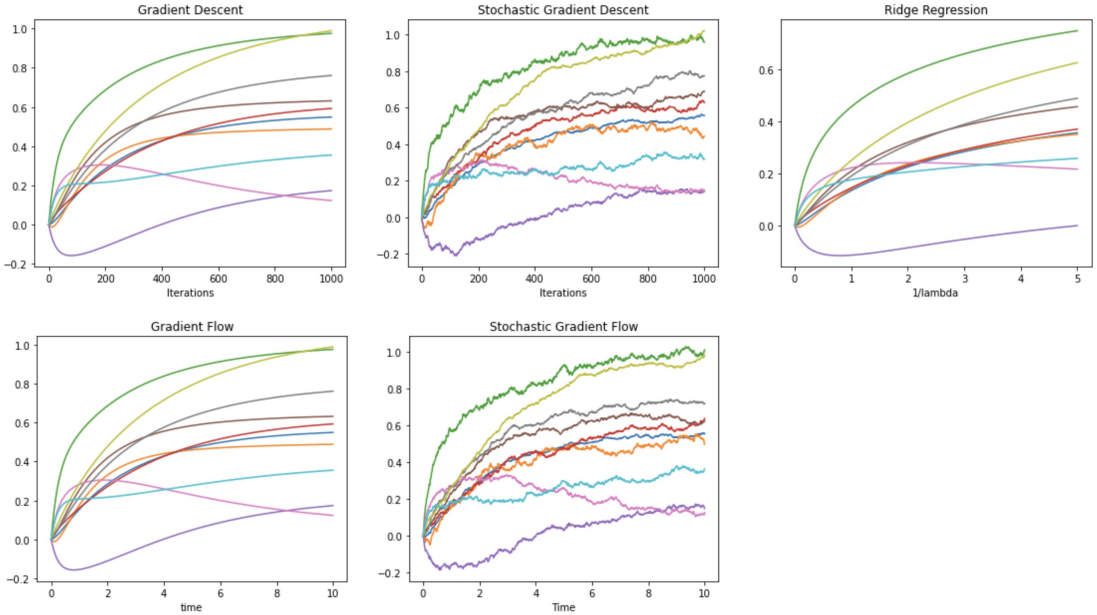
$$g^{\mathrm{ridge}}(s,\lambda)=\frac{s}{s+\lambda}, g^{\mathrm{gf}}(s,t)=1-\exp(-\mathrm{ts})$$

These two shrinkage maps agree at the extreme ends($\lambda=0$ and $t=\infty$, or $\lambda=\infty$ and $t=0$).

we can see a simple example which compares solution paths for different methods. We set $n=50$, $p=10$, $m=10(\mathrm{size\,of\,minibatch})$, $\varepsilon=0.01$, and generate data $X=\Sigma^{\frac{1}{2}}W$, where Σ is the covariance matrix whose diagonal entries are 1 and off-diagonals are 0.5. The ground truth is

$$y=X\beta_0+\eta, \eta\sim N(0,\sigma^2I)$$

where β_0 is a fixed random vector. We solve this linear regression problem using gradient descent, gradient flow, stochastic gradient descent, stochastic gradient flow, and ridge regression. The following are numerical results.



Here we use Euler discretization to solve stochastic gradient flow, we will talk more details later. We can see that the the above five solution paths are very similar. GD and GF are almost same since the this is a convex problem, gradient descent has nice convergence rate and stability. Then we can see that SGD can be a good approximation of SGF, only sublet difference. And ridge regression is very similar to each method, this gives us the intuition that ridge can be viewed as early stopping or GD/SGD has implicit regularization.