Data Mining Final

Hua Yao, UNI:hy2632

Problem 1: Reducing the variance of the Monte Carlo estimators [50 points]

Proposed estimator

Here we propose an estimator using antithetic sampling for variance reduction.

$$\hat{X}' = \frac{1}{m}\sum_{i=1}^{m}\frac{f(g_i^\top z)+f(-g_i^\top z)}{2}$$

Still $g_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$.

Proof: Unbiased

The unbiasedness is evident.

$$\begin{aligned}
E[\hat{X}'] &= E\left[\frac{1}{m}\sum_{i=1}^{m}\frac{f(g_i^\top z)+f(-g_i^\top z)}{2}\right] \\
&= \frac{1}{2}\left(E_{g\sim\mathcal{N}(0,I_d)}[f(g^\top z)] + E_{g\sim\mathcal{N}(0,I_d)}[f(-g^\top z)]\right) \\
&= \frac{1}{2}\left(E_{g\sim\mathcal{N}(0,I_d)}[f(g^\top z)] + E_{g\sim\mathcal{N}(0,I_d)}[f(g^\top z)]\right) && \because -g \sim \mathcal{N}(0, I_d) \\
&= E_{g\sim\mathcal{N}(0,I_d)}[f(g^\top z)] \\
&= E[X]
\end{aligned}$$

Proof: Strictly lower variance

Denote $x = g^\top z \in \mathbb{R}$, $X = f(x)$, x has a normal distribution parameterized by z, from the property of multivariate normal distribution.

$$x \sim N(0, \sum_{i=1}^{d}|z_i|)$$

Then the expectation w.r.t g can also be written as expectation w.r.t x. For an arbitrary function F,

$$E_g[F(g^\top z)] = E_x[F(x)]$$

The variance of estimators

$$\begin{aligned} Var(\hat{X}) &= Var\left[\frac{1}{m}\sum_{i=1}^{m} f(g_i^\top z)\right] \\ &= \frac{1}{m} Var\big[f(x)\big] \end{aligned}$$

$$\begin{aligned} Var(\hat{X}') &= Var\left[\frac{1}{m}\sum_{i=1}^{m} \frac{f(g_i^\top z) + f(-g_i^\top z)}{2}\right] \\ &= \frac{1}{m} Var\left[\frac{f(x) + f(-x)}{2}\right] \\ &= \frac{1}{4m}\Big(Var\big[f(x)\big] + Var\big[f(-x)\big] + 2Cov\big[f(x), f(-x)\big]\Big) \\ &= \frac{1}{2m}\Big(Var\big[f(x)\big] + Cov\big[f(x), f(-x)\big]\Big) \end{aligned}$$

To prove that $Var(\hat{X}') < Var(\hat{X})$, we only need to prove that $Cov\big[f(x), f(-x)\big] < Var\big[f(x)\big]$.

$$\begin{aligned} Cov\big[f(x), f(-x)\big] &= E\big[f(x)f(-x)\big] - E\big[f(x)\big]E\big[f(-x)\big] \\ &= E\big[f(x)f(-x)\big] - E\big[f(x)\big]^2 \end{aligned}$$

Here,

$$E\big[f(-x)\big] = E\big[f(x)\big]$$

because from above, $x \sim N(0, \sum_{i=1}^{d} |z_i|)$, $-x$ has the same distribution with x.

$$Var\big[f(x)\big] = E\big[f^2(x)\big] - E\big[f(x)\big]^2$$

$$\begin{aligned} Cov\big[f(x), f(-x)\big] - Var\big[f(x)\big] &= E\big[f(x)f(-x)\big] - E\big[f^2(x)\big] \\ &= E\big[f(x)f(-x) - f^2(x)\big] \\ &= \int_{-\infty}^{\infty} p(x)f(x)[f(-x) - f(x)]dx \end{aligned}$$

We know that f's taylor expansion has positive coefficients. Then it can be denoted as the sum of even and odd parts

$$f(x) = o(x) + e(x)$$

$p(x)$ of normal distribution is even.

$$
\begin{aligned}
\int_{-\infty}^{\infty} p(x)f(x)[f(-x)-f(x)]dx &= \int_{-\infty}^{\infty} p(x)[o(x)+e(x)][-2o(x)]dx \\
&= -2\int_{-\infty}^{\infty} p(x)[o^2(x)+o(x)e(x)]dx \\
&= -2\int_{-\infty}^{\infty} p(x)o^2(x)dx \qquad \text{drop the odd part in integration} \\
&< 0 \qquad \text{strictly less than 0 because o(x) has positive coefficients}
\end{aligned}
$$

$$\therefore Cov\big[f(x), f(-x)\big] < Var\big[f(x)\big]$$

$$\therefore Var(\hat{X}') < Var(\hat{X})$$

Problem 2: SVM with dynamically changing labels [30 points]

Known conditions

For given time t, from the derivation in the lagrangian program, we get

$$
\begin{aligned}
w_t &= \sum_{i=1}^{n} \alpha_t^{(i)} y_t^{(i)} x^{(i)} \\
&= (\alpha_t * y_t) \cdot x
\end{aligned}
$$

$w_t : (N,), \alpha_t : (N,), y_t : (N,), x : (N, d)$

And

$$\sum_{i=1}^{n} \alpha_t^{(i)} y_t^{(i)} = 0$$

Sufficient condition for the statement to hold

When $t \in \{0, ...N\}$, vectorized version

$$
\begin{aligned}
\begin{bmatrix} w_1 \\ ... \\ w_N \end{bmatrix} &= \begin{bmatrix} \alpha_1 * y_1 \\ ... \\ \alpha_N * y_N \end{bmatrix} \cdot x \\
&= \begin{bmatrix} \alpha_1^{(1)} y_1^{(1)}, ..., \alpha_1^{(N)} y_1^{(N)} \\ ... \\ \alpha_N^{(1)} y_N^{(1)}, ..., \alpha_N^{(N)} y_N^{(N)} \end{bmatrix}_{N \times N} \cdot \begin{bmatrix} x^{(1)} \\ ... \\ x^{(N)} \end{bmatrix}
\end{aligned}
$$

Denote the $N \times N$ matrix on the left as A. Then

$$w_t = A_t \cdot x$$

where

$$x = \begin{bmatrix} x^{(1)} \\ ... \\ x^{(N)} \end{bmatrix}_{N \times d}$$

if w_t is the linear combination of all other w vectors,

$$w_t = \sum_{j \in \{1,...N\}/t} c_j w_j$$

$$\Leftrightarrow A_t x = \sum_{j \in \{1,...N\}/t} c_j A_j x$$

$$\Leftrightarrow (A_t - \sum_{j \in \{1,...N\}/t} c_j A_j) x = 0$$

$\therefore$The **sufficient condition** that w_t can be the linear combination of all other w vectors is

$$A_t = \sum_{j \in \{1,...N\}/t} c_j A_j$$

This property is namely the linearly dependence of A's rows, i.e. **A is rank-deficient**.

Proof: A is rank-deficient

Using the fact that $\sum_{i=1}^{n} \alpha_t^{(i)} y_t^{(i)} = 0$, we can rewrite the rightmost entry of each row of A as negative the sum of all other entries in the row.

$$\begin{bmatrix} \alpha_1^{(1)} y_1^{(1)}, ..., \alpha_1^{(N-1)} y_1^{(N-1)}, -\sum_{i=1}^{N-1} \alpha_1^{(i)} y_1^{(i)} \\ ... \\ \alpha_N^{(1)} y_N^{(1)}, ..., \alpha_N^{(N-1)} y_N^{(N-1)}, -\sum_{i=1}^{N-1} \alpha_N^{(i)} y_N^{(i)} \end{bmatrix}$$

Obviously, the rightmost column of A is a linear combination of other columns, and the coefficients are all -1,

$$A_{:,N} = -\sum_{j=1}^{N-1} A_{:,j}$$

Then the column rank of A is strictly less than N, A is rank-deficient.

Therefore, there exists w_t which is a linear combination of all other vectors.

Problem 3: Gradient exploding/vanishing problems in neural network training [20 points]

Exploding/vanishing gradients and why it affects deep NN

In the simplest case where there are only linear transformations (without bias), denote the weight matrix between t-th and $(t+1)$-th as A.

$$\frac{\partial \ell}{\partial a_t} = A^T \frac{\partial \ell}{\partial a_{t+1}}$$

Then from the chain rule, partial derivative of the k-th hidden state and that of the loss satisfy:

$$\frac{\partial \ell}{\partial a_k} = \frac{\partial \ell}{\partial a_L} \cdot A_L^T A_{L-1}^T \ldots A_K^T$$

The product of matrices causes exploding/vanishing gradients. If the eigenvalues of these matrices are not restricted to be ± 1, the product of these matrices might have unbounded eigenvalues, especially when the number of As is large (which means the early layers suffer more from vanishing/exploding gradient problem).

Then when we do backpropagation using the formula above, the gradient for updating the parameters could be 0 (vanishing gradient) or ∞ (exploding gradient). The NN actually loses track of the gradient information of the loss function and we cannot update the parameters effectively.

Methods for handling this problem

Using orthogonal weight matrices in linear tranformations

The eigenvalues of orthogonal matrix are all ± 1. The product of orthogonal matrices is also orthogonal and has ± 1 eigenvalues and thus avoids exploding/vanishing gradient problem.

To parameterize the orthogonal matrix, we can use unrestricted parameterizations($|a_{ij}| \leq 1$), Givens Rotations matrices, or Riemannion optimization.

Using Residual Network

In many residual neural networks like ResNet, the shortcut skips some layers (especially linear layers). This help alleviate the problem of vanishing gradient.

Choice in activation function

Previously we mainly talked about the vanishing/exploding gradient problem in the setting of linear layers. In neural networks with activation functions, the chain rule formula also consists of the partial derivatives of these non-linear transformations.

Some activation functions might cause vanishing gradient by itself. For example, the partial derivative of sigmoid function is less than 0.25. Then such activations would make gradient smaller.

Therefore, ReLU and its variants (LeakyReLU, etc) is better in this respect.

Problem 4: ϵ-close estimators of kernels [40 points]

Denote the value of the kernel function $K(v,w)=X$. We use two estimators $\hat{X}_1,\hat{X}_2$ to approximate the exact value X.

When v,w fixed, denote the mean of two estimators $\hat{X}_1,\hat{X}_2$ as μ_1,μ_2. Since this is one state in the whole set S, we have

$$\|\mu_1-\mu_2\|\leq\delta$$

Then,

$$\begin{aligned}MSE(\hat{X}_1)&=E\big[(\hat{X}_1-\mu_1)^2+(\mu_1-X)^2\big]\\&=Var(\hat{X}_1)+bias(\hat{X}_1,X)^2\\&=E\big[\hat{X}_1^2\big]-\mu_1^2+E\big[(\mu_1-X)^2\big]\end{aligned}$$

$$\begin{aligned}MSE(\hat{X}_1)-MSE(\hat{X}_2)&=E\big[\hat{X}_1^2\big]-E\big[\hat{X}_2^2\big]-\mu_1^2+\mu_2^2+E\big[(\mu_1-X)^2\big]-E\big[(\mu_2-X)^2\big]\\&=E\big[\hat{X}_1^2\big]-E\big[\hat{X}_2^2\big]+2X(\mu_2-\mu_1)\end{aligned}$$

$$\begin{aligned}|MSE(\hat{X}_1)-MSE(\hat{X}_2)|&=|E\big[\hat{X}_1^2\big]-E\big[\hat{X}_2^2\big]+2X(\mu_2-\mu_1)|\\&\leq|E\big[\hat{X}_1^2\big]-E\big[\hat{X}_2^2\big]|+|2X(\mu_2-\mu_1)|\end{aligned}$$

Research the first part $|E\big[\hat{X}_1^2\big]-E\big[\hat{X}_2^2\big]|$. Since f is bounded by F,

$$\begin{aligned}|\hat{X}_i|&=|\frac{1}{m}f(G_iv)f(G_iw)|\\&\leq\frac{1}{m}|f(G_iv)||f(G_iw)|\\&\leq\frac{F^2}{m}\end{aligned}$$

$$\therefore 0\leq\hat{X}_i^2\leq\frac{F^4}{m^2}$$

$$
\begin{aligned}
E[\hat{X}_i^2] &= \int_{-\infty}^{\infty} f(\hat{X}_i)\hat{X}_i^2 d\hat{X}_i \\
&\leq \int_{-\infty}^{\infty} f(\hat{X}_i)\frac{F^4}{m^2} d\hat{X}_i \\
&= \frac{F^4}{m^2}\int_{-\infty}^{\infty} f(\hat{X}_i) d\hat{X}_i \\
&= \frac{F^4}{m^2}
\end{aligned}
$$

And similarly,

$$E[\hat{X_i^2}] \geq 0$$

$$\therefore |E\big[\hat{X_1^2}\big] - E\big[\hat{X_2^2}\big]| \leq \frac{2F^4}{m^2}$$

Then look at the second part $|2X(\mu_2 - \mu_1)|$.

$$
\begin{aligned}
|2X(\mu_2 - \mu_1)| &\leq |2X||\mu_2 - \mu_1| \\
&\leq |2X|\delta
\end{aligned}
$$

Here X is the exact kernel value $K(v, w)$. We already knew $|\hat{X}_i|$ is bounded by $\frac{F^2}{m}$. It goes without saying that, $|X|$ should also be bounded by a value, which is even lower than $\frac{F^2}{m}$. If not, $E[\hat{X}_i]$ cannot estimate X.

$$\therefore |X| \leq \frac{F^2}{m}$$

$$
\begin{aligned}
|2X(\mu_2 - \mu_1)| &\leq |2X|\delta \\
&\leq \frac{2F^2}{m}\delta
\end{aligned}
$$

To sum up,

$$
\begin{aligned}
|MSE(\hat{X}_1) - MSE(\hat{X}_2)| &= |E\big[\hat{X_1^2}\big] - E\big[\hat{X_2^2}\big] + 2X(\mu_2 - \mu_1)| \\
&\leq |E\big[\hat{X_1^2}\big] - E\big[\hat{X_2^2}\big]| + |2X(\mu_2 - \mu_1)| \\
&\leq \frac{2F^4}{m^2} + \frac{2F^2}{m}\delta = \epsilon
\end{aligned}
$$