Name: Arpit Gupta AG3418

COMS 4772 Homework Set 4

(1) You may use the fact that *expectation is a linear operator*.

(a) For a random variable X, let EX denote its expected value. Show that

$$E\left((X - EX)(X - EX)^T\right) = E(XX^T) - EX(EX)^T.$$

The quantity on the left hand side is the variance-covariance matrix for X, which we will call $V(X)$.

$$\begin{aligned}
&E\left(XX^T - (EX)X^T - X(EX)^T + EX(EX)^T\right) \\
&= E\left(XX^T - 2(EX)X^T + EX(EX)^T\right) \\
&= E\left(XX^T\right) - E\left(2(EX)X^T + EX(EX)^T\right) \\
&= E\left(XX^T\right) - E\left(2(EX)X^T\right) + E\left(EX(EX)^T\right) \\
&= E\left(XX^T\right) - 2(EX)E\left(X^T\right) + \left(EX(EX)^T\right) \\
&= E\left(XX^T\right) - 2(EX)\left(EX\right)^T + \left(EX(EX)^T\right) \\
&= E\left(XX^T\right) - EX\left(EX\right)^T \\
&= RHS \\
&HenceProved.
\end{aligned}$$

(b) Show that, for any (appropriately sized) matrix A we have

$$V(AX) = A(V(X))A^T.$$

$V(AX) = E(AX(AX)^T) - E(AX)(E(AX))^T$
$\implies V(AX) = E(AXX^TA^T) - E(AX)(E(AX))^T$
$\implies V(AX) = AE(XX^T)A^T - AE(X)(E(X)^TA^T$
$\implies V(AX) = A(E(XX^T) - E(X)(E(X)^T)A^T$
$\implies V(AX) = A(V(X))A^T$

Hence Proved .

(c) Show that

$$E(\|X\|^2) = \text{trace}(V(X)) + \|EX\|^2.$$

$\text{trace}(V(X)) + \|EX\|^2 = \text{trace}(E(XX^T) - E(X)E(X)^T) + \|EX\|^2$
$\implies \text{trace}(V(X)) + \|EX\|^2 = \text{trace}(E(XX^T)) - \text{trace}(E(X)E(X)^T) + \|EX\|^2$
$\implies \text{trace}(V(X)) + \|EX\|^2 = E(\text{trace}(XX^T)) - \text{trace}(E(X)E(X)^T) + \|EX\|^2$
Now, $\text{trace}(XX^T) = XX^T$ as XX^T is a scalar
$\implies \text{trace}(V(X)) + \|EX\|^2 = E(XX^T)$
$\implies \text{trace}(V(X)) + \|EX\|^2 = E(\|X\|^2)$

(d) Solve the stochastic optimization problem

$$\min_y E\|X - y\|_2^2,$$

where X is a random vector, and the expectation is taken with respect to X. What is the minimizer? What's the minimum value?

Answer :

$$\min_y(E\|X - y\|_2^2$$
$$= \min_y(E((X - y)(X - y))^T)$$
$$= \min_y(E(XX^T - 2Xy^T + yy^T)$$
$$= \min_y(E(XX^T) - 2y^T E(X) + yy^T)$$

Take gradeient of above equation and equate to zero.

$$\nabla(E(XX^T) - 2y^T E(X) + yy^T) = 0$$
$$\implies -2E(X) + 2y = 0$$
$$\implies y = E(X)$$

(2) Frobenius norm estimation. Suppose we want to estimate

$$\|A\|_F^2 = \text{trace}(A^T A)$$

of a large matrix A. One way to do this is to hit A by random vectors w, and then measure the resulting norm.

(a) Find a sufficient conditions on a random vector w that ensures

$$E\|Aw\|^2 = \|A\|_F^2.$$

Prove that your condition works.
Answer :

$$\|A\|_F^2 = \text{trace}(A^T A)$$
$$\implies \|A\|_F^2 = \text{trace}((A^T I A))$$
$$\implies \|A\|_F^2 = \text{trace}(A^T E(w^T w) A)$$

where w, is a random variable with E(W) = 0 , and Var(w) = 1

$$\implies \|A\|_F^2 = \text{trace}(A^T E(w^T w) A)$$
$$\implies \|A\|_F^2 = \text{trace}(E\left((wA)^T wA\right))$$

As trace is a linear operator,

$$\implies \|A\|_F^2 = E\left(\text{trace}((wA)^T wA)\right)$$
$$\implies \|A\|_F^2 = E\left(\text{trace}(\|Aw\|^2)\right)$$

Now, $\|Aw\|$ will, be a $1X1$ scalar , therefore it is same as its trace.

$$\implies \|A\|_F^2 = E\|Aw\|^2$$

Hence Proved.
And w, mjst be a random variable with E(W) = 0 , and Var(w) = 1

(b) What's a simple example of a distribution that satisfies the condition you derived above?

White Gaussian Noise is an example of w that will satisy above condition.

(c) Explain how you can put the relationship you found to practical use to estimate $\|A\|_F^2$ for a large A. In particular, you must explain how to estimate $\|A\|_F^2$ more or less accurately, depending on the need.

Answer :

We can take Expected value of $\|Aw\|^2$ by choosing different w multiple times, and averaging over the values of $\|Aw\|^2$ By increasing the number of times we sample w, we can achieve higer accuracy, as the sampling count approaches infinity, we will exactly match $\|A\|_F^2$

(d) Test out the idea in Matlab. Generate a random matrix A, maybe 500 x 1000. Compute its frobenius norm using `norm(A, 'fro')` command. Compare this to the result of your approach. Are they close? Is your approach faster?

Answer :

It is faster when number of times w is less than . As the number of times I sample w is increased, accuracy increases.

(3) Consider again the logistic regression problem. Included with this homework is the covtype dataset (500K examples, 54 features).
Consider again the logistic regression formulation:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(\tilde{x}_i^T \theta)) + \lambda \|\theta\|_2$$

where $\tilde{x}_i = -y_i x_i$ and you can take $\lambda = 0.01$ (small regularization).
Implement a stochastic gradient method for this problem.
Use the following options for step length:
(a) Pre-specified constant
(b) Decreasing with the rule $\alpha(k) \propto \frac{1}{k}$ (with some initialization)
(c) Decreasing with rule $\alpha(k) \propto \frac{1}{k^{0.6}}$ (with some initialization)

Divide covtype into two datasets, 90% training and 10% testing. Tune each of the three previous step size routines (i.e. adjust the constant or the constant initialization) until you are happy each one performs reasonably well. Make a graph showing the value of the *test likelihood* as a function of the iterates for each of the three strategies.

(4) (BONUS)

(a) Change the counting in the previous problem to be as a function of *effective passes through the data*, rather than iterations. For example, five iterations with batch size 1 should be no different than one iteration with batch size 5 in this metric.

(b) For the pre-specified constant step length strategy, compare test likelihood as a function of effective passes through the data for different random batch sizes, e.g. 1, 10, and 100.

(c) Again for pre-specified constant step length strategy, implement a growing batch size strategy, where the size of the batch increases with iterations. Can this strategy beat the fixed batch size strategy, with respect to effective passes through the data?