Yale University
Department of Statistics and Data Science

STATISTICS 365/565

Issued: 03/06/2021 **Due:** 03/22/2021

Notes: This hw covers exponential families, derivatives, vector functions, and optimization. Also, I will generally begin with examples in these problems.

Notation: $[k] = \{1, 2, \ldots, k\}$. For a matrix $A \in \mathbb{R}^{m \times n}$ we will let $A_{(i,\cdot)}$ denote the i^{th} row and $A_{(\cdot,j)}$ denote the j^{th} column. **Both will be treated as column vectors.**

1 Problem 1: Gradients

We will begin this problem with some examples. If you want to jump to the problems go to Section 1.2

1.1 Examples

Example 0: We have done this one before, but let's do it again. Suppose we have a vector valued function $f(v) = \langle v, w \rangle = v^T w = w^T v$ where $v, w \in \mathbb{R}^p$ Then, $\nabla f(v) = w$.

Proof. We begin by writing out $f(v) = v^T w = \sum_{i=1}^p v_{(i)} w_{(i)}$. The i^{th} coordinate of the gradient can be written as

$$\begin{aligned}
[\nabla f(v)]_{(i)} &= \frac{\partial f(v)}{\partial v_{(i)}} \\
&= \frac{\partial (\sum_{j=1}^p v_{(j)} w_{(j)})}{\partial v_{(i)}} \\
&= \sum_{j=1}^p \frac{\partial (v_{(j)} w_{(j)})}{\partial v_{(i)}} \quad \text{Linearity of derivative} \\
&= \sum_{j=1}^p \mathbb{1}(i = j) w_{(j)} \\
&= w_{(i)}
\end{aligned}$$

Therefore, $\nabla f(v) = w$ □

Example 1: Suppose that $f : \mathbb{R} \mapsto \mathbb{R}$ (again that means that $f(x) \in \mathbb{R}$ for $x \in \mathbb{R}$). Take $g(v) = f(v^T w)$. Then, $\nabla g(v) = \nabla_v f(v^T w) = f'(v^T w) w$. The notation ∇_v just means we are taking the gradient with respect to v since it is not clear in the previous equation. For this solution we simply just use the previous example and the chain rule for the file `notation.pdf` .

Example 2: For the next example consider the function $f(v) = v^T Av$ for an arbitrary matrix $A \in \mathbb{R}^{p \times p}$. We claim that $\nabla f(v) = Av + A^T v$

Proof. We again begin by writing out the i^{th} coordinate of the gradient

$$\begin{aligned}
[\nabla f(v)]_{(i)} &= \frac{\partial f(v)}{\partial v_{(i)}} \\
&= \frac{\partial(\sum_{k=1}^{p} \sum_{j=1}^{p} v_{(k)} v_{(j)} A_{(kj)})}{\partial v_{(i)}} \\
&= \sum_{k=1}^{p} \sum_{j=1}^{p} \frac{\partial(v_{(k)} v_{(j)} A_{(kj)})}{\partial v_{(i)}} \quad \text{Linearity of derivative} \\
&= \sum_{k=1}^{p} \sum_{j=1}^{p} v_{(j)} A_{(kj)} \mathbb{1}(i=k) + v_{(k)} A_{(kj)} \mathbb{1}(i=j) \quad \text{Product rule of derivative } (fg)' = f'g + fg' \\
&= \sum_{j=1}^{p} v_{(j)} A_{(ij)} + \sum_{k=1}^{p} v_{(k)} A_{(ki)} \\
&= [Av]_{(i)} + [A^T v]_{(i)}
\end{aligned}$$

If you have trouble with the inequalities please see the `notation.pdf` file under the section involving summations. Since the i^{th} coordinates match, our result clearly follows. □

Example 3: Let $A \in \mathbb{R}^{m \times n}$ be a matrix and let $C \in \mathbb{R}^{m \times n}$ be another matrix. Define the function $f(A) = \text{trace}(AC^T)$. Then, $\nabla_A f(A) = C$.

Proof. First we note that since f is a function of a matrix, its gradient is **also a matrix**. **This is NOT the Hessian. The Hessian of this function would be written as a fourth order**

tensor. Thus, we begin by analyzing its i, j coordinate

$$\begin{aligned}
[\nabla f(A)]_{(ij)} &= \frac{\partial f(A)}{\partial A_{(ij)}} \\
&= \frac{\partial \operatorname{trace}(AC^T)}{\partial A_{(ij)}} \\
&= \frac{\partial (\sum_{k=1}^m (AC^T)_{(kk)})}{\partial A_{(ij)}} \\
&= \frac{\partial (\sum_{k=1}^m \sum_{\ell=1}^n A_{(k\ell)} [C^T]_{(\ell k)})}{\partial A_{(ij)}} \\
&= \frac{\partial (\sum_{k=1}^m \sum_{\ell=1}^n A_{(k\ell)} C_{(k\ell)})}{\partial A_{(ij)}} \\
&= \sum_{k=1}^m \sum_{\ell=1}^n \frac{\partial (A_{(k\ell)} C_{(k\ell)})}{\partial A_{(ij)}} \\
&= \sum_{k=1}^m \sum_{\ell=1}^n \mathbb{1}(i=k, j=\ell) C_{(k\ell)} \\
&= C_{(ij)}
\end{aligned}$$

Thus we have established the desired result. □

The reader may have noticed another approach to the above problem that can leverage previous work. Namely, that $f(A) = \operatorname{trace}(AC^T) = \sum_{k=1}^m A^T_{(k,\cdot)} C_{(k,\cdot)}$.

1.1.1 A quick tool for computing gradients

Here is a method that often times acts as a shortcut for computing gradients. **You cannot use this method to solve the homework problems below. You can use it to check your work.**

If we recall the Taylor approximation then we know that for nice functions

$$f(v + \Delta) \approx f(v) + \langle \Delta, \nabla f(v) \rangle$$

where Δ is considered a small vector. We can use this to quickly read out derivatives. Let's go through all of the above examples.

Example 0: First we have $f(v) = v^T w$. Then we can write

$$\begin{aligned}
f(v + \Delta) &= (v + \Delta)^T w \\
&= v^T w + \Delta^T w \\
&= f(v) + \langle \Delta, w \rangle
\end{aligned}$$

Thus, we just look at whatever is in the inner-product with Δ and that's the gradient. In this case it is w which matches what we found above.

Example 1: Let $f : \mathbb{R} \mapsto \mathbb{R}$ and $v \in \mathbb{R}^p$. Take $g(v) = f(v^T w)$. Then, $\nabla_v g(v) = f'(v^T w)w$. To do this, just recall Taylor's approximation again, but this time for the univariate function f. Then $f(t+\delta) \approx f(t) + \delta f'(t)$ for δ small. We can write

$$\begin{aligned} g(v+\Delta) &= f((v+\Delta)^T w) \\ &= f(v^T w + \Delta^T w) \\ &\approx f(v^T w) + f'(v^T w)\Delta^T w \\ &= f(v^T w) + \Delta^T w f'(v^T w) \quad \text{since for } a, b \in \mathbb{R} \text{ we have } ab = ba \text{ and } \Delta^T w \in \mathbb{R}. \\ &= f(v^T w) + \langle \Delta, f'(v^T w)w \rangle \end{aligned}$$

We again see that the gradient is $f'(v^T w)w$, which matches our analysis above.

Example 2: Now $f(v) = v^T A v$. Again, we write

$$\begin{aligned} f(v+\Delta) &= (v+\Delta)^T A(v+\Delta) \\ &= v^T Av + \Delta^T Av + v^T A\Delta + \Delta^T A\Delta \\ &= f(v) + \Delta^T(Av + A^T v) + \text{ stuff} \\ &= f(v) + \langle \Delta, Av + A^T v \rangle + \text{ stuff} \end{aligned}$$

This time we have an extra term $\Delta^T A\Delta$. That term is considered higher order since it's quadratic, so we just move it into **stuff** and only consider the terms that are linear in Δ. Again, we read out the gradient as $Av + A^T v$, which matches.

Example 3: Now for the matrix problem. $f(A) = \text{trace}(AC^T)$. This time we let $\Delta \in \mathbb{R}^{m \times n}$ be a small matrix.

$$\begin{aligned} f(A+\Delta) &= \text{trace}((A+\Delta)C^T) \\ &= \text{trace}(AC^T) + \text{trace}(\Delta C^T) \\ &= f(A) + \langle \Delta, C \rangle \end{aligned}$$

So again, we see that C is the gradient and matches the computations from above.

1.2 Problems

All of the problems below should be solved with the same rigor as the above derivations. You cannot refer to the matrix cookbook or other resources nor can you use the "quick tool" from above. The derivations must come from basic multi-variable calculus. All vectors are p dimensional unless otherwise stated. All matrices are of commensurate dimension with the appropriate vectors.

a) Let $A, C \in \mathbb{R}^{m \times n}$. Show that $\text{trace}(AC^T) = \sum_{i=1}^m \sum_{j=1}^n A_{(ij)} C_{(ij)}$.

Remark: Recall the inner-product between two vectors v and w is $\langle v, w\rangle = \sum_{i=1}^{p} v_i w_i$. We can define an inner-product for matrices in exactly the same way. $\langle A, C\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{(ij)}C_{(ij)}$. You have just shown that that is equal to $\text{trace}(AC^T)$. And by the previous problem that is also equal to $\text{trace}(C^T A) = \langle C^T, A^T\rangle$. Also, obviously, $\langle A, C\rangle = \text{trace}(AC^T) = \text{trace}(CA^T) = \langle C, A\rangle$.

b) Compute the gradient of $f(v) = \sum_i v_{(i)}^3$

c) Let $X \in \mathbb{R}^{n\times p}$ be the matrix with i^{th} row x_i^T and $y_i \in \mathbb{R}$. Show that the gradient of $f(\beta) = \sum_i (x_i^T\beta - y_i)^3 = 3X^T(X\beta - y)^2$ where we take $(X\beta - y)^2$ to be the vector whose i^{th} entry is $(x_i^T\beta - y_i)^2$. **Hint:** Use the chain rule as discussed in `notation.pdf`. In general one can show that $X^T\epsilon = \sum_{i=1}^{n} x_i\epsilon_i$.

d) Compute the gradient $f(A) = \text{trace}(ACA^T)$ for $A \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{n\times n}$ an arbitrary matrix. **That means C is not necessarily symmetric.** Hint: Think about leveraging Example 3 from above as well as the product rule of derivative or using he chain rule.

2 Problem 2: Exponential families

Recall from class that an exponential family takes the form

$$p(y;\theta) = h(y)\exp\big(\langle\theta, T(y)\rangle) - A(\theta)\big)$$

where the parameters θ are known as the canonical parameters. Recall from lecture the derivation for the Bernoulli and Gaussian cases. The function $A(\theta)$ is known as the log-partition function. Its roll is to ensure that the distribution normalizes to one. Namely,

$$\begin{aligned}\int_y p(y;\theta) &= \int_y h(y)\exp\big(\langle\theta, T(y)\rangle) - A(\theta)\big)\\ &= \exp\big(-A(\theta)\big)\int_y h(y)\exp\big(\langle\theta, T(y)\rangle)\big)\\ &= 1\end{aligned}$$

Therefore, $\exp\big(A(\theta)\big) = \int_y h(y)\exp\big(\langle\theta, T(y)\rangle)\big)$. When we write these integrals we assume they are a sum when y is discrete. Note that $Z(\theta) = \int_y h(y)\exp\big(\langle\theta, T(y)\rangle)\big)$ is known as the partition function. Hence, $A(\theta)$ is the log-partition function.

Remark: For a wide class of important problems the log-partition function and partition function are difficult to compute, and a wide range of research goes into understanding their properties. Recall that we said that $\nabla A(\theta) = \mathbb{E}_\theta T(y)$. Where we take $\mathbb{E}_\theta$ to be the expectation when the underlying parameter is θ.

a) Consider the Bernoulli exponential family. Recall it can be written as

$$p(y) = \mathbb{1}(y \in \{0,1\})\exp(y\theta - \log(1+\exp(\theta)))$$

LOG IS NATURAL LOG Where $\theta \in \mathbb{R}$. Identify $h(y)$, $T(y)$, and $A(\theta)$.

b) Now consider the Gaussian distribution $N(\mu, \sigma^2)$. In this problem we only care about the mean parameter μ. What are $h(y)$, $T(y)$, and $A(\mu)$? Note that σ^2 will appear in your solutions. Where we are taking $\theta = \mu$ in this case.

c) Again, for the above Bernoulli. Compute $A^{'}(\theta)$ and confirm that $A^{'}(\theta) = \mathbb{E}_\theta y$. (Recall that $A^{'}(\theta)$ is just the derivative of $A(\theta)$ evaluated at θ. We use derivative since for the Bernoulli θ is just a scalar and A is a scalar valued function.)

d) Suppose that we observe n i.i.d. samples from the Bernoulli distribution. Compute the maximum likelihood estimate for θ by computing the gradient of the log-likelihood, setting it to zero, and solving for θ. Recall that the likelihood can be written as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} p(y_i; \theta) \\ &= \prod_{i=1}^{n} \mathbb{1}(y_i \in \{0, 1\}) \exp(y_i\theta - \log(1 + \exp(\theta))) \end{aligned}$$

REMEMBER THAT WE ARE ASSUMING THE y_i ARE OBSERVED, BUT THEY ARE STILL RANDOM VARIABLES. SO THERE IS A PROBABILITY OF OBSERVING THEM.

We don't need the indicator part anymore because we obviously have that $y_i \in \{0, 1\}$ already. So dropping that and taking logs we get

$$\log L(\theta) = \sum_{i=1}^{n} y_i\theta - \log(1 + \exp(\theta))$$

e) For the general exponential family. Prove that MLE estimate $\widehat{\theta}$ for θ given n i.i.d. samples from the exponential family distribution satisfies the equality

$$\nabla A(\widehat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} T(y_i)$$

3 Problem 3: Generalized Linear Models and Gradient descent

Recall the generalized linear model. For an example (x_i, y_i) we have that

$$p(y_i; x_i, \theta^*) = h(y_i) \exp(y_i \langle x_i, \theta^* \rangle - A(\langle x_i, \theta^* \rangle)$$

a) Suppose that we have n i.i.d. copies of (x_i, y_i). Verify that the negative log-likelihood can be written as the following for a candidate choice parameter θ.

$$\sum_{i=1}^{n} A(\langle x_i, \theta \rangle) - y_i \langle x_i, \theta \rangle - \log(h(y_i))$$

b) Verify that the gradient with respect to θ can be written as

$$\sum_{i=1}^{n} x_i (A^{'}(\langle x_i, \theta \rangle) - y_i)$$

Hint: Just look at Example 1 above.

Remark: We can see that $A'(\langle x_i, \theta \rangle) - y_i$ acts as a sort of error. To make this explicit recall that $A'(s)$ is the expected value of the response in our exponential family for parameter s. Therefore, $A'(\langle x_i, \theta \rangle) = \mathbb{E}[y_i | x_i, \theta]$ is the expected response if we assume θ is the true parameter and given x_i. Thus, the deviation between the true observed y_i and this expected value is exactly the error for a general linear model. Check what it is for the Gaussian case to convince yourself.

c) For logistic regression compute $A^{'}(\langle x_i, \theta \rangle) - y_i$. Recall that $A(t) = \log(1 + \exp(t))$.

Next, we will look at stochastic gradient descent for GLMs. We pick a uniform random example $J \in [n]$ and update

$$\theta_k = \theta_{k-1} - \eta_k x_{J_k}((A^{'}(\langle x_{J_k}, \theta_{k-1} \rangle) - y_{J_k})$$

We assume that $\|x_i\|_2 > 0$ for all i.

d) Consider the case for linear regression. That is take $A(s) = \frac{s^2}{2}$ and $T(y) = y$. Compute the SGD update.

e) Let $\eta_k = \frac{1}{10\|x_{J_k}\|_2^2}$. Show that

$$|A^{'}(\langle x_{J_k}, \theta_k \rangle) - y_{J_k}| < |A^{'}(\langle x_{J_k}, \theta_{k-1} \rangle) - y_{J_k}|$$

Remark: Note that this is not how one selects the step-size. The point of the problem is to just show that SGD is trying to correct an individual example at each round.