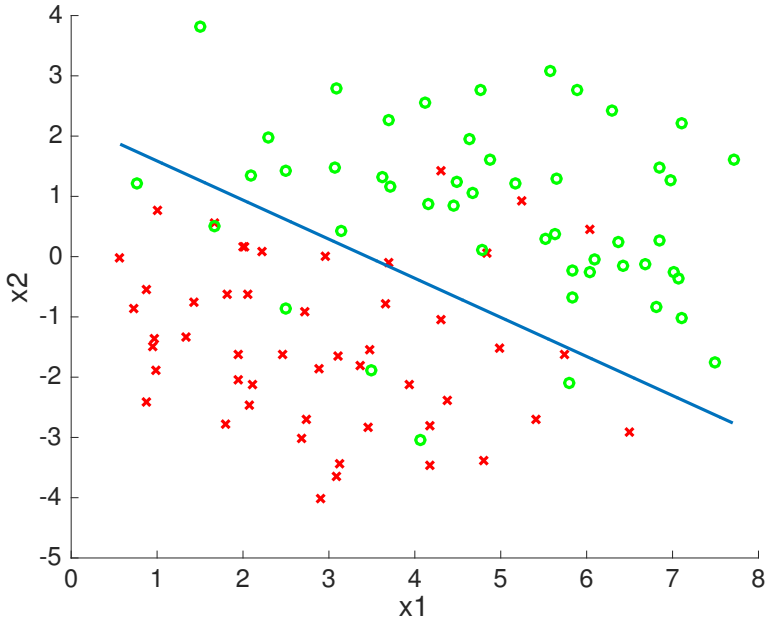

Figure 1: Separating hyperplane for logistic regression (question 1c).

(a) [5 points] Consider the Poisson distribution parameterized by λ:

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

Show that the Poisson distribution is in the exponential family, and clearly state what are $b(y)$, η, $T(y)$, and $a(\eta)$.

Answer: Rewrite the distribution function as:

$$\begin{aligned} p(y;\lambda) &= \frac{e^{-\lambda}e^{y\log\lambda}}{y!} \\ &= \frac{1}{y!}\exp(y\log\lambda - \lambda) \end{aligned}$$

Comparing with the standard form for the exponential family:

$$\begin{aligned} b(y) &= \frac{1}{y!} \\ \eta &= \log\lambda \\ T(y) &= y \\ a(\eta) &= e^{\eta} \end{aligned}$$

(b) [3 points] Consider performing regression using a GLM model with a Poisson response variable. What is the canonical response function for the family? (You may use the fact that a Poisson random variable with parameter λ has mean λ.)

Answer: The canonical response function for the GLM model will be:

$$\begin{aligned} g(\eta) &= E[y;\eta] \\ &= \lambda \\ &= e^{\eta} \end{aligned}$$

(c) [7 points] For a training set $\{(x^{(i)}, y^{(i)});\ i = 1, \ldots, m\}$, let the log-likelihood of an example be $\log p(y^{(i)}|x^{(i)};\theta)$. By taking the derivative of the log-likelihood with respect to θ_j, derive the stochastic gradient ascent rule for learning using a GLM model with Poisson responses y and the canonical response function.

Answer: The log-likelihood of an example $(x^{(i)}, y^{(i)})$ is defined as $\ell(\theta) = \log p(y^{(i)}|x^{(i)};\theta)$. To derive the stochastic gradient ascent rule, use the results in part (a) and the standard GLM assumption that $\eta = \theta^T x$.

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \theta_j} &= \frac{\partial \log p(y^{(i)}|x^{(i)};\theta)}{\partial \theta_j} \\ &= \frac{\partial \log \left(\frac{1}{y^{(i)}!} \exp(\eta^T y^{(i)} - e^{\eta}) \right)}{\partial \theta_j} \\ &= \frac{\partial \log \left(\exp((\theta^T x^{(i)})^T y^{(i)} - e^{\theta^T x^{(i)}}) \right)}{\partial \theta_j} + \frac{\partial \log \left(\frac{1}{y^{(i)}!} \right)}{\partial \theta_j} \\ &= \frac{\partial \left((\theta^T x^{(i)})^T y^{(i)} - e^{\theta^T x^{(i)}} \right)}{\partial \theta_j} \\ &= \frac{\partial \left((\sum_k \theta_k x_k^{(i)}) y^{(i)} - e^{\sum_k \theta_k x_k^{(i)}} \right)}{\partial \theta_j} \\ &= x_j^{(i)} y^{(i)} - e^{\sum_k \theta_k x_k^{(i)}} x_j^{(i)} \\ &= (y^{(i)} - e^{\theta^T x^{(i)}}) x_j^{(i)} \end{aligned}$$

Thus the stochastic gradient ascent update rule should be:

$$\theta_j := \theta_j + \alpha \frac{\partial \ell(\theta)}{\partial \theta_j}$$

which reduces here to:

$$\theta_j := \theta_j + \alpha (y^{(i)} - e^{\theta^T x}) x_j^{(i)}$$

(d) **[3 extra credit points]** Consider using GLM with a response variable from any member of the exponential family in which $T(y) = y$, and the canonical response function $h(x)$ for the family. Show that stochastic gradient ascent on the log-likelihood $\log p(\vec{y}|X;\theta)$ results in the update rule $\theta_i := \theta_i - \alpha(h(x) - y)x_i$.

Answer: As in the previous part, consider the derivative of the likelihood of a training example

(x, y) with respect to the parameter θ_j:

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta_j} &= \frac{\partial \log p(y|x;\theta)}{\partial \theta_j} \\
&= \frac{\partial \log \left(b(y)\exp(\eta^T y - a(\eta))\right)}{\partial \theta_j} \\
&= \frac{\partial \left(\eta^T y - a(\eta))\right)}{\partial \theta_j} \\
&= x_j y - \frac{\partial a(\eta)}{\partial \eta} x_j \\
&= \left(y - \frac{\partial a(\eta)}{\partial \eta}\right) x_j
\end{aligned}$$

Thus, it only remains to show that $\frac{\partial a(\eta)}{\partial \eta} = h(x) = E[y|x;\theta]$. To prove this consider the fact that $p(y|x;\theta)$ is a probability distribution and must thus sum to 1.

$$\begin{aligned}
\int_y p(y|x;\theta)dy &= 1 \\
\int_y b(y)\exp(\eta^T y - a(\eta))dy &= 1 \\
\int_y b(y)\exp(\eta^T y)dy &= \exp(a(\eta))
\end{aligned}$$

Differentiating both sides with respect to η:

$$\begin{aligned}
\int_y b(y) y \exp(\eta^T y)dy &= \exp(a(\eta))\frac{\partial a(\eta)}{\partial \eta} \\
\frac{\partial a(\eta)}{\partial \eta} &= \int_y b(y) y \exp(\eta^T y - a(\eta))dy \\
&= \int_y y p(y|x;\theta)dy \\
&= E[y|x;\theta]
\end{aligned}$$

where the last step follows from the definition of the (conditional) expectation of a random variable. Substituting this into the expression for $\frac{\partial \ell(\theta)}{\partial \theta_j}$ gives the required gradient ascent update rule.

3. **[15 points] Gaussian discriminant analysis**

Suppose we are given a dataset $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$ consisting of m independent examples, where $x^{(i)} \in \mathbb{R}^n$ are n-dimensional vectors, and $y^{(i)} \in \{-1, 1\}$. We will model the joint

distribution of (x, y) according to:

$$\begin{aligned} p(y) &= \begin{cases} \phi & \text{if } y = 1 \\ 1-\phi & \text{if } y = -1 \end{cases} \\ p(x|y=-1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})\right) \\ p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) \end{aligned}$$

Here, the parameters of our model are ϕ, Σ, μ_{-1} and μ_1. (Note that while there're two different mean vectors μ_{-1} and μ_1, there's only one covariance matrix Σ.)

(a) [5 points] Suppose we have already fit ϕ, Σ, μ_{-1} and μ_1, and now want to make a prediction at some new query point x. Show that the posterior distribution of the label at x takes the form of a logistic function, and can be written

$$p(y \mid x; \phi, \Sigma, \mu_{-1}, \mu_1) = \frac{1}{1+\exp(-y(\theta^T x + \theta_0))},$$

where $\theta \in \mathbb{R}^n$ and the bias term $\theta_0 \in \mathbb{R}$ are some appropriate functions of $\phi, \Sigma, \mu_{-1}, \mu_1$. (Note: the term θ_0 corresponds to introducing an extra coordinate $x_0^{(i)} = 1$, as we did in class.)

Answer: For shorthand, we let $\mathcal{H} = \{\phi, \Sigma, \mu_{-1}, \mu_1\}$ denote the parameters for the problem. Since the given formulae are conditioned on y, use Bayes rule to get:

$$\begin{aligned} p(y=1|x;\phi,\Sigma,\mu_{-1},\mu_1) &= \frac{p(x|y=1;\phi,\Sigma,\mu_{-1},\mu_1)p(y=1;\phi,\Sigma,\mu_{-1},\mu_1)}{p(x;\phi,\Sigma,\mu_{-1},\mu_1)} \\ &= \frac{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H}) + p(x|y=-1;\mathcal{H})p(y=-1;\mathcal{H})} \\ &= \frac{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)\phi}{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)\phi + \exp\left(-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})\right)(1-\phi)} \\ &= \frac{1}{1+\frac{1-\phi}{\phi}\exp\left(-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1}) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)} \\ &= \frac{1}{1+\exp\left(\log(\frac{1-\phi}{\phi}) - \frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1}) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}. \end{aligned}$$

Now, we expand and rearrange the difference of quadratic terms in the preceding expression, finding that

$$\begin{aligned} &(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1}) - (x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \\ &= x^T\Sigma^{-1}x - \mu_{-1}^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_{-1} + \mu_{-1}^T\Sigma^{-1}\mu_{-1} - x^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}\mu_1 \\ &= -2\mu_{-1}^T\Sigma^{-1}x + \mu_{-1}^T\Sigma^{-1}\mu_{-1} + 2\mu_1^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}\mu_1 \\ &= 2(\mu_1-\mu_{-1})^T\Sigma^{-1}x + \mu_{-1}^T\Sigma^{-1}\mu_{-1} - \mu_1^T\Sigma^{-1}\mu_1. \end{aligned}$$

Thus, we have

$$p(y=1 \mid x; \mathcal{H}) = \frac{1}{1+\exp\left(\log\frac{1-\phi}{\phi} + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1} + (\mu_{-1}-\mu_1)^T\Sigma^{-1}x\right)}.$$

and setting

$$\theta = \Sigma^{-1}(\mu_1 - \mu_{-1}) \text{ and } \theta_0 = \frac{1}{2}(\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \mu_1^T\Sigma^{-1}\mu_1) - \log\frac{1-\phi}{\phi}$$

gives that

$$p(y \mid x; \phi, \Sigma, \mu_{-1}, \mu_1) = \frac{1}{1+\exp(-y(\theta^T x + \theta_0))}.$$

(b) [10 points] For this part of the problem only, you may assume n (the dimension of x) is 1, so that $\Sigma = [\sigma^2]$ is just a real number, and likewise the determinant of Σ is given by $|\Sigma| = \sigma^2$. Given the dataset, we claim that the maximum likelihood estimates of the parameters are given by

$$\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^m 1\{y^{(i)} = 1\} \\
\mu_{-1} &= \frac{\sum_{i=1}^m 1\{y^{(i)} = -1\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = -1\}} \\
\mu_1 &= \frac{\sum_{i=1}^m 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T
\end{aligned}$$

The log-likelihood of the data is

$$\begin{aligned}
\ell(\phi, \mu_{-1}, \mu_1, \Sigma) &= \log\prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi, \mu_{-1}, \mu_1, \Sigma) \\
&= \log\prod_{i=1}^m p(x^{(i)}|y^{(i)}; \mu_{-1}, \mu_1, \Sigma)p(y^{(i)}; \phi).
\end{aligned}$$

By maximizing ℓ with respect to the four parameters, prove that the maximum likelihood estimates of ϕ, μ_{-1}, μ_1, and Σ are indeed as given in the formulas above. (You may assume that there is at least one positive and one negative example, so that the denominators in the definitions of μ_{-1} and μ_1 above are non-zero.)

Answer: The derivation follows from the more general one for the next part.

(c) **[3 extra credit points]** Without assuming that $n = 1$, show that the maximum likelihood estimates of ϕ, μ_{-1}, μ_1, and Σ are as given in the formulas in part (b). [Note: If you're fairly sure that you have the answer to this part right, you don't have to do part (b), since that's just a special case.]

Answer: First, derive the expression for the log-likelihood of the training data:

$$\begin{aligned}
\ell(\phi, \mu_{-1}, \mu_1, \Sigma) &= \log\prod_{i=1}^m p(x^{(i)}|y^{(i)}; \mu_{-1}, \mu_1, \Sigma)p(y^{(i)}; \phi) \\
&= \sum_{i=1}^m \log p(x^{(i)}|y^{(i)}; \mu_{-1}, \mu_1, \Sigma) + \sum_{i=1}^m \log p(y^{(i)}; \phi) \\
&\simeq \sum_{i=1}^m \big[\frac{1}{2}\log\frac{1}{|\Sigma|} - \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}}) + y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)\big]
\end{aligned}$$

where constant terms indepedent of the parameters have been ignored in the last expression.

Now, the likelihood is maximized by setting the derivative (or gradient) with respect to each of the parameters to zero.

$$\begin{aligned}\frac{\partial \ell}{\partial \phi} &= \sum_{i=1}^{m} \big[\frac{y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi}\big] \\ &= \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}{\phi} - \frac{m - \sum_{i=1}^{m} 1\{y^{(i)}=1\}}{1-\phi}\end{aligned}$$

Setting this equal to zero and solving for ϕ gives the maximum likelihood estimate.

For μ_{-1}, take the gradient of the log-likelihood, and then use the same kinds of tricks as were used to analytically solve the linear regression problem.

$$\begin{aligned}\nabla_{\mu_{-1}} \ell &= -\frac{1}{2} \sum_{i:y^{(i)}=-1} \nabla_{\mu_{-1}} (x^{(i)} - \mu_{-1})^T \Sigma^{-1} (x^{(i)} - \mu_{-1}) \\ &= -\frac{1}{2} \sum_{i:y^{(i)}=-1} \nabla_{\mu_{-1}} \big[\mu_{-1}^T \Sigma^{-1} \mu_{-1} - {x^{(i)}}^T \Sigma^{-1} \mu_{-1} - \mu_{-1}^T \Sigma^{-1} x^{(i)}\big] \\ &= -\frac{1}{2} \sum_{i:y^{(i)}=-1} \nabla_{\mu_{-1}} tr\big[\mu_{-1}^T \Sigma^{-1} \mu_{-1} - {x^{(i)}}^T \Sigma^{-1} \mu_{-1} - \mu_{-1}^T \Sigma^{-1} x^{(i)}\big] \\ &= -\frac{1}{2} \sum_{i:y^{(i)}=-1} \big[2\Sigma^{-1}\mu_{-1} - 2\Sigma^{-1} x^{(i)}\big]\end{aligned}$$

The last step uses matrix calculus identities (specifically, those given in page 8 of the lecture notes), and also the fact that Σ (and thus Σ^{-1}) is symmetric.

Setting this gradient to zero gives the maximum likelihood estimate for μ_{-1}. The derivation for μ_1 is similar to the one above.

For Σ, we find the gradient with respect to $S = \Sigma^{-1}$ rather than Σ just to simplify the derivation (note that $|S| = \frac{1}{|\Sigma|}$). You should convince yourself that the maximum likelihood estimate S_m found in this way would correspond to the actual maximum likelihood estimate Σ_m as $S_m^{-1} = \Sigma_m$.

$$\begin{aligned}\nabla_S \ell &= \sum_{i=1}^{m} \nabla_S \big[\frac{1}{2}\log|S| - \frac{1}{2} \underbrace{(x^{(i)} - \mu_{y^{(i)}})^T}_{b_i^T} S \underbrace{(x^{(i)} - \mu_{y^{(i)}})}_{b_i}\big] \\ &= \sum_{i=1}^{m} \big[\frac{1}{2|S|}\nabla_S |S| - \frac{1}{2}\nabla_S b_i^T S b_i\big]\end{aligned}$$

But, we have the following identities:

$$\nabla_S |S| = |S|(S^{-1})^T$$

$$\nabla_S b_i^T S b_i = \nabla_S tr\left(b_i^T S b_i\right) = \nabla_S tr\left(S b_i b_i^T\right) = b_i b_i^T$$

In the above, we again used matrix calculus identities, and also the commutatitivity of the trace operator for square matrices. Putting these into the original equation, we get:

$$\begin{aligned}\nabla_S \ell &= \sum_{i=1}^{m} \big[\frac{1}{2}S^{-1} - \frac{1}{2}b_i b_i^T\big] \\ &= \frac{1}{2}\sum_{i=1}^{m} \big[\Sigma - b_i b_i^T\big]\end{aligned}$$

Setting this to zero gives the required maximum likelihood estimate for Σ.

4. **[10 points] Linear invariance of optimization algorithms**

Consider using an iterative optimization algorithm (such as Newton's method, or gradient descent) to minimize some continuously differentiable function $f(x)$. Suppose we initialize the algorithm at $x^{(0)} = \vec{0}$. When the algorithm is run, it will produce a value of $x \in \mathbb{R}^n$ for each iteration: $x^{(1)}, x^{(2)}, \ldots$.

Now, let some non-singular square matrix $A \in \mathbb{R}^{n\times n}$ be given, and define a new function $g(z) = f(Az)$. Consider using the same iterative optimization algorithm to optimize g (with initialization $z^{(0)} = \vec{0}$). If the values $z^{(1)}, z^{(2)}, \ldots$ produced by this method necessarily satisfy $z^{(i)} = A^{-1}x^{(i)}$ for all i, we say this optimization algorithm is **invariant to linear reparameterizations**.

(a) [7 points] Show that Newton's method (applied to find the minimum of a function) is invariant to linear reparameterizations. Note that since $z^{(0)} = \vec{0} = A^{-1}x^{(0)}$, it is sufficient to show that if Newton's method applied to $f(x)$ updates $x^{(i)}$ to $x^{(i+1)}$, then Newton's method applied to $g(z)$ will update $z^{(i)} = A^{-1}x^{(i)}$ to $z^{(i+1)} = A^{-1}x^{(i+1)}$.[3]

Answer: Let $g(z) = f(Az)$. We need to find $\nabla_z g(z)$ and its Hessian $\nabla_z^2 g(z)$.

By the chain rule:

$$\begin{aligned}\frac{\partial g(z)}{\partial z_i} &= \sum_{k=1}^{n} \frac{\partial f(Az)}{\partial (Az)_k}\frac{\partial (Az)_k}{\partial z_i} &(1)\\ &= \sum_{k=1}^{n} \frac{\partial f(Az)}{\partial (Az)_k} A_{ki} &(2)\\ &= \sum_{k=1}^{n} \frac{\partial f(Az)}{\partial x_k} A_{ki} &(3)\end{aligned}$$

Notice that the above is the same as :

$$\frac{\partial g(z)}{\partial z_i} = A_{\bullet i}^{\top}\nabla_x f(Az) \quad (4)$$

where $A_{\bullet i}$ is the i'th column of A. Then,

$$\nabla_z g(z) = A^{\top}\nabla_x f(Az) \quad (5)$$

[3] Note that for this problem, you must explicitly prove any matrix calculus identities that you wish to use that are not given in the lecture notes.

where $\nabla_x f(Az)$ is $\nabla_x f(\cdot)$ evaluated at Az.

Now we want to find the Hessian $\nabla_z^2 g(z)$.

$$\begin{aligned}\frac{\partial^2 g(z)}{\partial z_i \partial z_j} &= \frac{\partial}{\partial z_j} \sum_{k=1}^{n} \frac{\partial f(Az)}{\partial (Az)_k} A_{ki} &(6)\\ &= \sum_l \sum_k \frac{\partial^2 f(Az)}{\partial x_l \partial x_k} A_{ki} A_{lj} &(7)\end{aligned}$$

If we let $H_f(y)$ denote the Hessian of $f(\cdot)$ evaluated at some point y, and let $H_g(y)$ be the Hessian of $g(\cdot)$ evaluated at some point y, we have from the previous equation that:

$$H_g(z) = A^\top H_f(Az) A \tag{8}$$

We can now put this together and find the update rule for Newton's method on the function $f(Ax)$:

$$\begin{aligned}z^{(i+1)} &= z^{(i)} - H_g(z^{(i)})^{-1} \nabla_z g(z^{(i)}) &(9)\\ &= z^{(i)} - (A^\top H_f(Az^{(i)})A)^{-1} A^\top \nabla_x f(Az^{(i)}) &(10)\\ &= z^{(i)} - A^{-1} H_f(Az^{(i)})^{-1} (A^\top)^{-1} A^\top \nabla_x f(Az^{(i)}) &(11)\\ &= z^{(i)} - A^{-1} H_f(Az^{(i)})^{-1} \nabla_x f(Az^{(i)}) &(12)\end{aligned}$$

Now we have the update rule for $z^{(i+1)}$, we just need to verify that $z^{(i+1)} = A^{-1}x^{(i+1)}$ or equivalently that $Az^{(i+1)} = x^{(i+1)}$. From Eqn. (12) we have

$$\begin{aligned}Az^{(i+1)} &= A\left(z^{(i)} - A^{-1} H_f(Az^{(i)})^{-1} \nabla_x f(Az^{(i)})\right) &(13)\\ &= Az^{(i)} - H_f(Az^{(i)})^{-1} \nabla_x f(Az^{(i)}) &(14)\\ &= x^{(i)} - H_f(x^{(i)})^{-1} \nabla_x f(x^{(i)}) &(15)\\ &= x^{(i+1)}, &(16)\end{aligned}$$

where we used in order: Eqn. (12); rewriting terms; the inductive assumption $x^{(i)} = Az^{(i)}$; the update rule $x^{(i+1)} = x^{(i)} - H_f(x^{(i)})^{-1} \nabla_x f(x^{(i)})$.

(b) [3 points] Is gradient descent invariant to linear reparameterizations? Justify your answer.

Answer:

No. Using the notation from above, gradient descent on $g(z)$ results in the following update rule:

$$z^{(i+1)} = z^{(i)} - \alpha A^\top \nabla_x f(Az^{(i)}). \tag{17}$$

The update rule for $x^{(i+1)}$ is given by

$$x^{(i+1)} = x^{(i)} - \alpha \nabla_x f(x^{(i)}). \tag{18}$$

The invariance holds if and only if $x^{(i+1)} = Az^{(i+1)}$ given $x^{(i)} = Az^{(i)}$. However we have

$$\begin{aligned} Az^{(i+1)} &= Az^{(i)} - \alpha AA^\top \nabla_x f(Az^{(i)}) && (19) \\ &= x^{(i)} - \alpha AA^\top \nabla_x f(x^{(i)}). && (20) \end{aligned}$$

The two expressions in Eqn. (18) and Eqn. (20) are not necessarily equal ($AA^T = I$ requires that A be an orthogonal matrix), and thus gradient descent is not invariant to linear reparameterizations.

5. **[35 points] Regression for denoising quasar spectra**[4]

Introduction. In this problem, we will apply a supervised learning technique to estimate the light spectrum of *quasars*. Quasars are luminous distant galactic nuclei that are so bright, their light overwhelms that of stars in their galaxies. Understanding properties of the spectrum of light emitted by a quasar is useful for a number of tasks: first, a number of quasar properties can be estimated from the spectra, and second, properties of the regions of the universe through which the light passes can also be evaluated (for example, we can estimate the density of neutral and ionized particles in the universe, which helps cosmologists understand the evolution and fundamental laws governing its structure). The *light spectrum* is a curve that relates the light's intensity (formally, lumens per square meter), or *luminous flux*, to its wavelength. Figure 2 shows an example of a quasar light spectrum, where the wavelengths are measured in Angstroms (Å), where 1Å$= 10^{-10}$ meters.

The Lyman-α wavelength is a wavelength beyond which intervening particles at most negligibly interfere with light emitted from the quasar. (Interference generally occurs when a photon is absorbed by a neutral hydrogen atom, which only occurs for certain wavelengths of light.) For wavelengths greater than this Lyman-α wavelength, the observed light spectrum f_{obs} can be modeled as a smooth spectrum f plus noise:

$$f_{\text{obs}}(\lambda) = f(\lambda) + \text{noise}(\lambda)$$

For wavelengths below the Lyman-α wavelength, a region of the spectrum known as the Lyman-α forest, intervening matter causes attenuation of the observed signal. As light emitted by the quasar travels through regions of the universe richer in neutral hydrogen, some of it is absorbed, which we model as

$$f_{\text{obs}}(\lambda) = \text{absorption}(\lambda) \cdot f(\lambda) + \text{noise}(\lambda)$$

Astrophysicists and cosmologists wish to understand the absorption function, which gives information about the Lyman-α forest, and hence the distribution of neutral hydrogen in otherwise unreachable regions of the universe. This gives clues toward the formation and evolution of the universe. Thus, it is our goal to estimate the spectrum f of an observed quasar.

Getting the data. We will be using data generated from the Hubble Space Telescope Faint Object Spectrograph (HST-FOS), Spectra of Active Galactic Nuclei and Quasars.[5] We have provided two comma-separated data files located at:

- Training set: `http://cs229.stanford.edu/ps/ps1/quasar_train.csv`
- Test set: `http://cs229.stanford.edu/ps/ps1/quasar_test.csv`

[4] Ciollaro, Mattia, et al. "Functional regression for quasar spectra." arXiv:1404.3168 (2014).
[5] `https://hea-www.harvard.edu/FOSAGN/`