CS 189 Introduction to Machine Learning
Spring 2018 HW8

Your self-grade URL is http://eecs189.org/self_grade?question_ids=1_1,1_2,2_1,2_2,2_3,2_4,2_5,2_6,2_7,2_8,3_1,3_2,3_3,3_4,3_5,4_1,4_2,4_3,4_4,5.

2 SGD on OLS

In this problem, we carefully walk through the key ingredients of a proof for SGD convergence for the specific example of a loss function which we are very familiar with: Ordinary Least Squares with positive definite matrix $\mathbf{X}^\top\mathbf{X}$. In particular we show that in this case, even though gradient descent converges to the optimal solution of OLS, with small enough constant stepsize, SGD is not guaranteed to do so! We then show that in contrast, when applying SGD with decaying step, as outlined in lecture, it does converge to the same optimum as gradient descent.

This phenomenon that constant stepsize gets stuck and decaying stepsizes are the way out is analogous to common observations in practical machine learning problems, where decreasing the learning rate using heuristic learning rate schedules helps to decrease the training error.

Recall that the ordinary least squares problem can be written as:

$$\min_{\mathbf{w}} f(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{n}||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 = \min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{w} - \mathbf{y}_i)^2 = \min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

where $f_i(\mathbf{w}) := (\mathbf{x}_i^\top\mathbf{w} - \mathbf{y}_i)^2$. Here $\mathbf{x}_i^\top$ is the ith row of matrix $\mathbf{X}$ and f_i is the loss of the training example i. We implement SGD as follows:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha_t \nabla f_{i_t}(\mathbf{w}^t).$$

where i_t is uniformly sampled from all samples $\{1, 2, \cdots, n\}$ (and is independently drawn for each iteration t). Let's define the short hand $G_t = G(\mathbf{w}^t) = \nabla f_{i_t}(\mathbf{w}^t)$ to represent the random gradient at step t. We are interested in how $\mathbf{w}^t$ approaches the optimal solution $\mathbf{w}^* := \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$. One way to characterize this is to monitor the squared distance of the iterate to the optimum, i.e. $||\mathbf{w}^t - \mathbf{w}^*||_2^2$. Throughout, we will assume that for fixed $\mathbf{w}$, the following bound holds for the squared norm of the stochastic gradient:

$$\mathbb{E}||G(\mathbf{w})||_2^2 \leq M_g^2||\mathbf{w} - \mathbf{w}^*||_2^2 + B^2, \tag{1}$$

where M_g and B are constants dependent on the model and loss function f. We will find concrete values of them later in this problem for specific examples. (Notice that in lecture we assumed M_g to be zero, which is too restrictive even for the most basic least squares loss.)

Problem outline: Parts (a)-(c) help to derive a recursive formula of the form $\mathbb{E}||\mathbf{w}^{t+1} - \mathbf{w}^*||_2^2 \leq \gamma\mathbb{E}||\mathbf{w}^t - \mathbf{w}^*||_2^2 + \tilde{\gamma}$. In part (d) we show that we achieve linear (also called geometric) convergence of $\mathbf{w}^t$ to the optimum $\mathbf{w}^*$ for SGD with constant stepsize when $f(\mathbf{w}^*) = 0$. That means we can write $\mathbb{E}||\mathbf{w}^t - \mathbf{w}^*||_2^2 \leq \gamma^t||\mathbf{w}^0 - \mathbf{w}^*||_2^2$ for some $\gamma < 1$. In part (e) we explore what happens with the iterates of SGD with constant stepsize if instead $f(\mathbf{w}^*) > 0$. In part (f) we visualize how the convergence bounds translate to actual training error decrease for OLS using SGD and GD.

The bonus parts (g)-(h) are for those students who "love math" (according to the midterm preliminary questions), where you are asked to prove that SGD with decaying stepsize converges as $1/t$ for "nice" functions using induction.

(a) **Show the following relation between the $t+1$-step error and the t-step error:** $||\mathbf{w}^{t+1} - \mathbf{w}^*||_2^2 = ||\mathbf{w}^t - \mathbf{w}^*||_2^2 - 2\alpha_t\langle G_t, \mathbf{w}^t - \mathbf{w}^*\rangle + \alpha_t^2||G_t||_2^2$

Solution:

$$||\mathbf{w}^{t+1} - \mathbf{w}^*||_2^2 = ||\mathbf{w}^t - \alpha_t G_t - \mathbf{w}^*||_2^2 \tag{2}$$
$$= \langle(\mathbf{w}^t - \mathbf{w}^*) - \alpha_t G_t, (\mathbf{w}^t - \mathbf{w}^*) - \alpha_t G_t\rangle \tag{3}$$
$$= ||\mathbf{w}^t - \mathbf{w}^*||_2^2 - 2\alpha_t\langle G_t, \mathbf{w}^t - \mathbf{w}^*\rangle + \alpha_t^2||G_t||_2^2 \tag{4}$$

(b) *Here we prove the key relation which underlies the success of stochastic gradient methods. This is where we use that stochastic gradients are unbiased!* Since we have stochastic gradients, we want to make guarantees in terms of expectations. For notational convenience let us define the short hand $\Delta_t := \mathbb{E}||\mathbf{w}^t - \mathbf{w}^*||_2^2$, the *average squared error*, where the expectation is taken over all the random indices drawn by the stochastic gradient method up to time t. We want to show a clean relation between Δ_t and Δ_{t+1} when the stochastic gradient satisfies (1).

Remind yourself which random indices $\mathbf{w}^t$ depends on and use the unbiasedness of the stochastic gradient to **show that**

$$\mathbb{E}[\langle G(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\star\rangle] = \mathbb{E}[\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\star\rangle]. \tag{5}$$

Now use equality (5) **and part (a) to prove that**

$$\Delta_{t+1} \leq (1 + \alpha_t^2 M_g^2)\Delta_t - 2\alpha_t\mathbb{E}\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle + \alpha_t^2 B^2. \tag{6}$$

Hint 1: You may take the law of iterated expectation (also called tower property) as given, i.e.

$$\mathbb{E}[\langle G(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\star\rangle] = \mathbb{E}_{i_1,\ldots,i_{t-1}}[\mathbb{E}_{i_t}[\langle G(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\star\rangle | i_1, \ldots, i_{t-1}]].$$

See Discussion for the derivation.

Hint 2: Unbiased stochastic gradient means that $\mathbb{E}_{i_t}\nabla f_{i_t}(\mathbf{w}) = \nabla f(\mathbf{w})$ for any $\mathbf{w}$ independent of the random index (note that this is the case here), where the expectation is with respect to the random index i_t at time t. Refer to the Discussion for brief intro on conditional expectations.

Solution: Three steps are needed here:

$$\begin{aligned}\mathbb{E}[\langle G(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\star\rangle] &= \mathbb{E}_{i_1,\ldots,i_{t-1}}[\mathbb{E}_{i_t}[\langle G(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\star\rangle | i_1, \ldots, i_{t-1}]] \\ &= \mathbb{E}_{i_1,\ldots,i_{t-1}}[\langle \mathbb{E}_{i_t}[G(\mathbf{w}^t)|i_1, \ldots, i_{t-1}], \mathbf{w}^t - \mathbf{w}^\star\rangle] \\ &= \mathbb{E}[\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^\star\rangle]\end{aligned}$$

First is tower property (given), second is the fact that $\mathbf{w}^t$ does NOT depend on i_t (which is the random index to compute $\mathbf{w}^{t+1}$), third is unbiasedness.

Explaining the second equality: Now we know that taken $i_1, \ldots, i_{k-1}$ as given (that is what conditioning is doing) the only randomness lies in the the index we sample in step k. However the gradient is unbiased when taking the expectation over the indices as shown in (b). Therefore

$$\mathbb{E}_{i_k}[G(\mathbf{w}^k)|i_1, \ldots, i_{k-1}] = \nabla f(\mathbf{w}^k)$$

Taking expectations w.r.t. the randomness in $\mathbf{w}^t$ and apply the assumption above, Then the average squared error satisfies the following recursion:

$$\begin{aligned}\Delta_{t+1} &= \mathbb{E}||\mathbf{w}^{t+1} - \mathbf{w}^*||_2^2 && (7)\\ &= \mathbb{E}||\mathbf{w}^t - \mathbf{w}^*||_2^2 - 2\alpha_t\mathbb{E}\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle + \alpha_t^2\mathbb{E}||G_t||_2^2 && (8)\\ &\leq \Delta_t - 2\alpha_t\mathbb{E}\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle + \alpha_t^2 M_g^2\mathbb{E}||\mathbf{w}_t - \mathbf{w}^*||_2^2 + \alpha_t^2 B^2 && (9)\\ &= (1 + \alpha_t^2 M_g^2)\Delta_t - 2\alpha_t\mathbb{E}\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle + \alpha_t^2 B^2 && (10)\end{aligned}$$

(c) *In this step we see that we need for our analysis that our loss function is "nice" (strongly convex to be precise), which holds for OLS when $\lambda_{min}(\mathbf{X}^\top\mathbf{X}) > 0$. That is, all the guarantees for SGD we discuss in this problem only hold for such "nice" functions. In general, at least convexity is needed for any convergence proof.* **Using inequality** (6)**, now show that $\Delta_{t+1} \leq (1 + \alpha_t^2 M_g^2 - 2\alpha_t m)\Delta_t + \alpha_t^2 B^2$, where we assume that the minimum eigenvalue of matrix $\mathbf{X}^\top\mathbf{X}$ denoted by m is positive, i.e. $m := \frac{2\lambda_{min}(\mathbf{X}^\top\mathbf{X})}{n} > 0$.**

Hint: Use the fact that $\nabla f(\mathbf{w}^*) = \mathbf{0}$, and hence

$$\langle \nabla f(\mathbf{w}^t), \mathbf{w} - \mathbf{w}^*\rangle = \langle \nabla f(\mathbf{w}^t) - \nabla f(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^*\rangle.$$

Solution: Let's look at the term $\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle$. Since $\mathbf{w}^*$ is the optimum, we have $\nabla f(\mathbf{w}^*) = 0$, thus

$$\begin{aligned}&\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle && (11)\\ =&\langle \nabla f(\mathbf{w}^t) - \nabla f(\mathbf{w}^*), \mathbf{w}^t - \mathbf{w}^*\rangle && (12)\\ =&\frac{2}{n}\langle (\mathbf{X}^\top\mathbf{X}\mathbf{w}_t - \mathbf{X}^\top\mathbf{y}) - (\mathbf{X}^\top\mathbf{X}\mathbf{w}^* - \mathbf{X}^\top\mathbf{y}), \mathbf{w}^t - \mathbf{w}^*\rangle && (13)\\ =&\frac{2}{n}\langle \mathbf{X}^\top\mathbf{X}(\mathbf{w}_t - \mathbf{w}^*), \mathbf{w}^t - \mathbf{w}^*\rangle && (14)\end{aligned}$$

$$= \frac{2}{n}(\mathbf{w}_t - \mathbf{w}^*)^\top(\mathbf{X}^\top\mathbf{X})(\mathbf{w}^t - \mathbf{w}^*) \tag{15}$$

$$\geq \frac{2}{n}\lambda_{min}(\mathbf{X}^\top\mathbf{X})||\mathbf{w}_t - \mathbf{w}^*||_2^2 = m\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 \tag{16}$$

Thus we have:

$$\Delta_{t+1} \leq (1 + \alpha_t^2 M_g^2)\Delta_t - 2\alpha_t \mathbb{E}\langle \nabla f(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle + \alpha_t^2 B^2 \tag{17}$$

$$\leq (1 + \alpha_t^2 M_g^2 - 2\alpha_t m)\Delta_t + \alpha_t^2 B^2 \tag{18}$$

(d) We now examine how close we can get to the optimal solution $\mathbf{w}^*$ using SGD with **constant stepsize**, i.e. $\alpha_t = \alpha$ for all t in two different scenarios. First, for this question we assume that $\mathbf{X}\mathbf{w}^* = \mathbf{y}$, or in other words, that the minimum loss is 0. **Show that inequality** (1) **holds with** $B = 0$ **and** $M_g^2 = \max_i 4||\mathbf{x}_i||_2^4$ **in this case. Find the optimum learning rate and show that** $\Delta_t \leq (1 - \frac{m^2}{M_g^2})^t \Delta_0$**.** This means that with $B = 0$ we have linear (geometric) convergence!

Hint: You may want to use that for any matrix $\mathbf{A}$ and vector $\mathbf{w}$ it holds that $\|\mathbf{A}\mathbf{w}\|_2^2 \leq \lambda_{max}(\mathbf{A}^\top\mathbf{A})\|\mathbf{w}\|_2^2$ and the same bound applies when taking the expectation on both sides.

Solution:

$$\begin{aligned}
&\mathbb{E}||G(\mathbf{w})||_2^2 \\
=&4\mathbb{E}||\mathbf{x}_i\mathbf{x}_i^\top\mathbf{w} - \mathbf{x}_i\mathbf{y}_i||_2^2 \\
=&4\mathbb{E}||\mathbf{x}_i\mathbf{x}_i^\top\mathbf{w} - \mathbf{x}_i\mathbf{x}_i^\top\mathbf{w}^*||_2^2 \\
=&4\mathbb{E}||\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{w} - \mathbf{w}^*)||_2^2 \\
\leq&\mathbb{E}\lambda_{max}^2(\mathbf{x}_i\mathbf{x}_i^\top)||(\mathbf{w} - \mathbf{w}^*)||_2^2 \\
=&4\mathbb{E}||\mathbf{x}_i||_2^4||\mathbf{w} - \mathbf{w}^*||_2^2 \\
\leq&4\max_i ||\mathbf{x}_i||_2^4||\mathbf{w} - \mathbf{w}^*||_2^2
\end{aligned}$$

Thus in this case, $M_g = \max_i 4||\mathbf{x}_i||_2^4$ and $B = 0$.

When $B = 0$, and set $\alpha_t = \frac{m}{M_g^2}$ by setting derivative to zero, we have $\Delta_t \leq (1 - \frac{m^2}{M_g^2})^t\Delta_0$. This implies linear (geometric) convergence, like standard gradient descent, except with a different convergence rate.

(e) Instead of assuming that there exists a $\mathbf{w}^*$ that satisfies $\mathbf{X}\mathbf{w}^* = \mathbf{y}$ exactly, we now consider the case where $f(\mathbf{w}^*) = \frac{1}{n}\|\mathbf{X}\mathbf{w}^* - y\|^2 > 0$. One can show that in this case $\mathbb{E}||G(\mathbf{w})||_2^2 \leq M_g^2||\mathbf{w} - \mathbf{w}^*||_2^2 + B^2$ holds for some M_g and B, where $M_g \neq 0$ and $B \neq 0$. Suppose that you are given the constants M_g and B and the stepsize is still constant, i.e. $\alpha_t = \alpha$ for all t.

We define the short hand notation $\gamma := 1 + \alpha^2 M_g^2 - 2\alpha m$ and assume α is large enough such that $\gamma < 1$. **Using part (c), prove that** $\Delta_t \leq \gamma^t\Delta_0 + \frac{\alpha B^2}{2m - \alpha M_g^2}$**.** We now want to interpret this bound. **Does it guarantee convergence** $\mathbf{w}^t \to \mathbf{w}^\star$ **when** t **goes to infinity? Explain.**

Hint: You may find the following inequality helpful: $\sum_{t=0}^{n} \gamma^t \leq \frac{1}{1-\gamma}$ for $0 < \gamma < 1$.

Solution: Let's set $\beta = \alpha^2 B^2$ for simplicity of notation. From (c), we have $\Delta_{t+1} \leq \gamma \Delta_t + \beta$ and thus $\Delta_{t+1} - \frac{\beta}{1-\gamma} \leq \gamma(\Delta_t - \frac{\beta}{1-\gamma})$ (since simplifying the latter inequality gives you the first one). Applying this for t steps, we have $\Delta_t - \frac{\beta}{1-\gamma} \leq \gamma^t(\Delta_0 - \frac{\beta}{1-\gamma})$ and thus

$$\begin{aligned}
\Delta_t &\leq \gamma^t(\Delta_0 - \frac{\beta}{1-\gamma}) + \frac{\beta}{1-\gamma} \\
&= \gamma^t \Delta_0 + \frac{\beta}{1-\gamma}(1-\gamma^t) \\
&\leq \gamma^t \Delta_0 + \frac{\alpha B^2}{-\alpha M_g^2 + 2m}
\end{aligned}$$

where the last term follows because $\gamma^t < 1$ and $\frac{\beta}{1-\gamma} = \frac{\alpha^2 B^2}{-\alpha^2 M_g^2 + 2\alpha m} = \frac{\alpha B^2}{-\alpha M_g^2 + 2m}$.

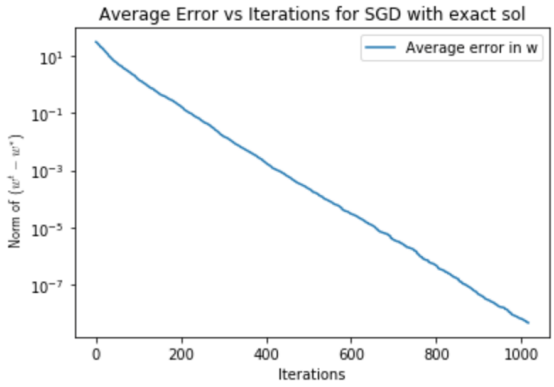
We can see that this bound does not guarantee Δ_t goes to zero, when t goes to infinity. This is what happens in real world: constant-step-size stochastic gradient descent does not converge to the optimum solution if $B \neq 0$, instead it bounces around a ball near the optimum.

(f) We now want use simulations to compare the first order methods for solving OLS for the cases in (d) and (e). In particular, we want to plot the estimation error $||\mathbf{w}^t - \mathbf{w}^*||_2$ against the gradient descent step in the following 3 scenarios:

- When there exists an exact solution $\mathbf{X}\mathbf{w}^* = \mathbf{y}$, plot the errors when you are using SGD and a constant learning rate.
- When there does not exist an exact solution $\mathbf{X}\mathbf{w}^* = \mathbf{y}$, plot the errors when you are using SGD and a constant learning rate. Also plot the errors for GD and the same constant learning rate.
- When there does not exist an exact solution $\mathbf{X}\mathbf{w}^* = \mathbf{y}$, plot the errors when you are using SGD and a decaying learning rate.

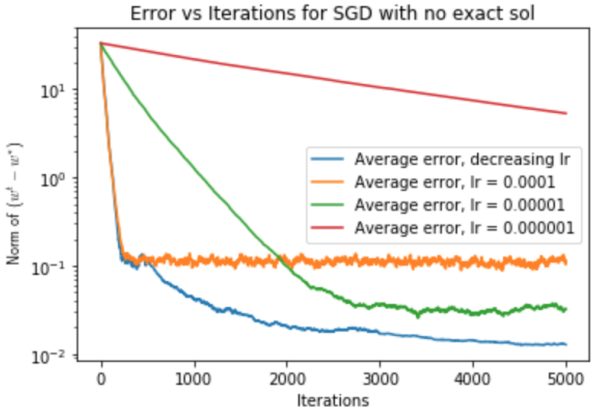
Using the attached starter code, **implement the gradient updates** in the function `sgd()` , while all the plotting functions are already there. **Show the 3 plots you obtain using the starter code. Report the average squared error computed in the starter code. What's your conclusion?**

Solution:



The MSE is 4.6e-09.

For lr = 0.0001, the final average error is 0.10. For lr = 0.00001, the final average error is 0.033. For lr = 0.000001, the final average error is 5.32, because it has not converged after 5000 iterations. If it had converged, it would reach a lower error. The GD final error is 3.04e-5. The iteration in the plot is how many gradient steps taken. Note that each GD step is n times more computationally expensive than a SGD step, where n is the number of data points involved.



The SGD with decaying learning rate has an error of 0.013 at iteration 5000 and it would continue to decrease if we run more steps.

In conclusion, when there is an exact solution, SGD with constant learning rate can find the optimal solution. When there is no exact solution, SGD with constant learning rate will approach a place close to the optimal, and the closeness is related to the learning rate. However, when the learning rate is too small, it will take a long time to converge. GD with a constant learning rate can still converge to the optimal. Lastly, SGD with decaying learning rate can converge to the optimal and faster than SGD with a small enough learning rate.

(g) (Bonus) **Prove that there is some constant S and k_0 such that if the learning rate is decaying, the error Δ_k satisfies $\Delta_k \leq \frac{S}{k+k_0}$.** In this part, prove for the case of $M_g = 0$ and $B \neq 0$.

Hint: induction is your friend.

Solution: Via the steps above we readily have the recursive inequality

$$\Delta_{k+1} \leq (1 - 2\alpha_k m)\Delta_k + \alpha_k^2 B^2 \tag{19}$$

For simplicity we refer to $\kappa = 1 - 2\alpha_k m$ as the contraction coefficient (remember contraction from last homework). We require $\kappa > 0$ and hence $\alpha_k \leq \frac{1}{2m}$.

Let us now find k_0, S such that $\Delta_k = \frac{S}{k+k_0}$ holds. In particular, we find the k_0, S such that an inductive proof goes through.

First, for the induction base case we require $d_0 \leq \frac{S}{k_0}$. For the induction step, assuming that the upper bound is correct for k, $\Delta_k = \frac{S}{k+k_0}$, we require

$$\begin{aligned} &\Delta_{k+1} \leq (1 - 2\alpha_k m)\frac{S}{k+k_0} + \alpha_k^2 B^2 \leq \frac{S}{k+k_0+1} \\ \iff\ &0 \geq \alpha_k^2 B^2 - \alpha_k \frac{2mS}{k+k_0} + \frac{S}{(k+k_0+1)(k+k_0)} \end{aligned}$$

to hold.

Solving the quadratic inequality, one can see that it is sufficient that $\alpha_k \leq \frac{mS}{B^2(k+k_0)}$ if $m^2S^2(k+k_0+1) - B^2S(k+k_0) \geq 0$, that is $S > \frac{B^2}{m^2}(1 - \frac{1}{k+k_0+1})$ for all k. Note that this holds if both $\alpha_k \leq \frac{1}{m(k+k_0)}$ and $S > \frac{B^2}{m^2}$. Additionally requiring $k_0 \geq 2$ also ensures $\alpha_k \leq \frac{1}{2m}$ for all k so that finally, the constraints containing S, k_0 boil down to just requiring

- $k_0 \geq 2$
- $d_0 \leq \frac{S}{k_0}$
- and $S \geq \frac{B^2}{m^2}$.

Choosing $k_0 = 2$ we consolidate the two constraints on S by taking the sum $S = 2d_0 + \frac{B^2}{m^2}$. As a consequence, $\alpha_k \leq \frac{1}{m(k+2)}$ works and $\Delta_k \leq \frac{2m^2d_0+B^2}{m^2(k+2)}$ is proven.

(h) (Bonus) **Prove that the same conclusion holds for the case of $M_g \neq 0$ and $B \neq 0$, for some different constant M_g and B.**

Hint: this general case can be reduced to the case above.

Solution: In the case of $M_g \neq 0$, we want to essentially reduce the problem to the case of $M_g = 0$ so that we are able to use the same induction argument.

$$\Delta_{k+1} \leq (1 - 2\alpha_k(m - \alpha_k M_g^2/2))\Delta_k + \alpha_k^2 B^2 \tag{20}$$

Let us now introduce γ and add the requirement $\alpha_k \leq \gamma = \frac{m}{M_g^2}$. Then

$$m - \alpha_k M_g^2/2 \geq m - \gamma M_g^2/2 = \frac{m}{2}$$

Then a sufficient constraint for (20) is

$$(1 - 2\alpha_k \tilde{m})\frac{S}{k + k_0} + \alpha_k^2 B^2 \leq \frac{S}{k + k_0 + 1},$$

and we are facing to prove the same inequality as in (g) just with an additoinal constraint $\alpha_k \leq \gamma$. The entirety of all constraints now reads

- $\alpha_k \leq \frac{1}{m(k+k_0)}$ for all k
- $\alpha_k \leq \frac{m}{M_g^2}$ for all k
- $d_0 \leq \frac{S}{k_0}$
- $k_0 \geq 2$
- and $S \geq \frac{B^2}{\tilde{m}^2}$.

We first solve for k_0

$$\frac{1}{m(k + k_0)} \leq \frac{m}{M_g^2}$$

which, together with requiring $k_0 \geq 2$ legitimates the choice $k_0 = 2(\frac{M_g^2}{m^2} + 1)$. Then we proceed as above and find that $S = k_0 d_0 + \frac{4B^2}{m^2}$ (per condition on induction base) satisfies the conditions for $\alpha_k \leq \frac{1}{m(k+k_0)}$ and thus we have proven that

$$\Delta_k \leq \frac{S}{k + k_0}. \tag{21}$$

3 Gradient Descent Framework

In HW1, you modeled the classification of digit numbers of the MNIST dataset as a linear regression problem and solved it using its closed-form solution. In this homework, you will model it better by using classification models such as logistic regression and neural networks, and solve it using stochastic gradient descent. The goal of this problem is to show the power of modern machine learning frameworks such as TensorFlow and PyTorch, and how they can solve a wide range of problems. TensorFlow is the recommended framework in this homework and we also provide the starter kit in PyTorch.

(a) The starter code contains an implementation of linear regression for the MNIST dataset to classify 10 digits. Let $\mathbf{x}_i$ be the feature vector and $\mathbf{y}_i$ be a one-hot vector encoding of its class, i.e., $y_{i,j} = 1$ if and only if ith images is in class j and $y_{i,j} = 0$ if not. In order to use linear regression to solve a multi-class classification, we will use the *one-vs-all* strategy here. In particular, for each j we will fit a linear regression model $(\mathbf{w}_j, b_j)$ to minimize $\sum_i (\mathbf{w}_j{}^\top \mathbf{x}_i + b_j - y_{i,j})_2^2$. Then for any image $\mathbf{x}$, the prediction of its class will be $\arg\max_j (\mathbf{w}_j{}^\top \mathbf{x} + b_j)$.

Read the implementation and run it with batch size equal to 50, 100, and 200. **Attach the "epoch vs validation accuracy" plot. Report the running time for all the batch sizes. Explain the potential benefits and drawbacks of using small batch size, i.e., SGD vs GD.**

Solution:

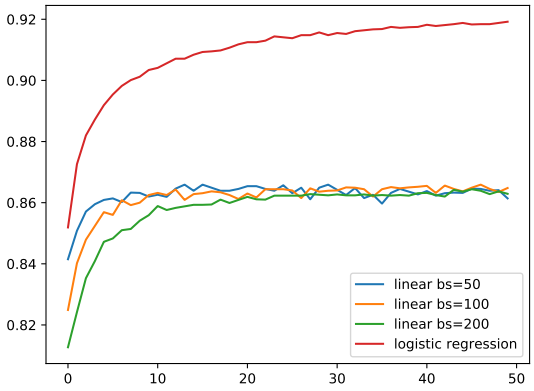


Figure 1: Epoch vs Accuracy

Figure 1 shows the plot. The running time for 50 epoches on my laptop is 51s, 28s, and 21s, respectively. The larger the batch size, the less time it takes to finish a constant number of epoches. The is because for larger batch size, there is less overhead due to the vectorization. On the other hand, when we use large batch size, we can see that it converges slower in practice. So, there is a tradeoff here.

(b) **Implement the `train_logistic` function to do the multi-class logistic regression using softmax function. Attach the plot of the "epoch vs validation accuracy" curve.**. The loss function of the multi-class logistic regression is

$$\ell = -\sum_{i=1}^{n} \log \left[\mathbf{softmax} \left(\mathbf{W}\mathbf{x}_i + \mathbf{b} \right)^\top \mathbf{y}_i \right] \tag{22}$$

where the softmax function $\mathbf{softmax} : \mathbb{R}^d \to \mathbb{R}^d$ is defined as

$$\text{softmax}(\mathbf{z})_j = \frac{\exp(z_j)}{\sum_k \exp(z_k)} = \frac{\exp(z_j - z')}{\sum_k \exp(z_k - z')}. \tag{23}$$

Here $z' = \max_j z_j$. The expression on the right is a numerical stable formula to compute the softmax. You may NOT use any functions in `tf.nn.*`, `tf.losses.*`, and `torch.nn.*` for all the parts of this problem.

Solution:

See Figure 1.

Note that the curve in the plot depends on the learning rate. If you are using a different learning rate, you will get different plot. There is some confusion on whether we should average or sum the error for training. The answer is that they are the same because the loss function are differed by a constant scale, so that you can just multiply or divide the learning rate by your batch size to get the exact same plot.

(c) Copy your code from `train_logistic` to `train_nn` and add an additional $\tanh$ nonlinear layer to it. You loss function should be something like

$$\ell = -\sum_i \log \left[\mathbf{softmax}\left(\mathbf{W}^{(2)} \mathbf{tanh}\left(\mathbf{W}^{(1)} \mathbf{x}_i + \mathbf{b}^{(1)} \right) + \mathbf{b}^{(2)} \right)^\top \mathbf{y}_i \right]. \tag{24}$$

Attach the plot of the "epoch vs validation accuracy" curve. You have the freedom but **are NOT required to** to choose the hyper-parameters to get the best performance.

Solution:

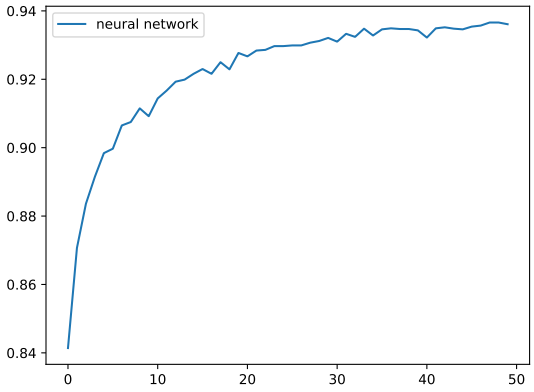


Figure 2: Epoch vs Accuracy (NN)

See Figure 2.

(d) In our previous lecture, we learned how to solve the total least squares by doing the SVD decomposition. Now we will try to model the problem from the probabilistic perspective and solve it using stochastic gradient descent. Let the probabilistic model be

$$y_i - z_y = \mathbf{w}^\top (\mathbf{x}_i - \mathbf{z}_\mathbf{x}) \tag{25}$$

where $\mathbf{x}_i$ and y_i is the observed data, $y_i - z_y$ is y_{true}, $\mathbf{x}_i - z_\mathbf{x}$ is x_{true}, $z_y \sim \mathcal{N}(0,1)$, and $z_{\mathbf{x},j} \sim \mathcal{N}(0,1)$ for all j. **Prove that the log-likelihood for the model in Equation** (25) **is**

$$\sum_{i=1}^n \log P_\mathbf{w}(y_i \mid \mathbf{x}_i) = C - \frac{n}{2} \log\left(\|\mathbf{w}\|_2^2 + 1\right) - \frac{1}{2\left(\|\mathbf{w}\|_2^2 + 1\right)} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i)^2, \tag{26}$$

where n is the number of samples, and C is a constant that is not related to $\mathbf{w}$, y, and $\mathbf{x}$. *Note that the maximum likelihood estimation of this probabilistic model is not the same as the SVD solution of TLS as we did in lecture.*

Solution: We have

$$y_i = \mathbf{w}^\top (\mathbf{x}_i - \mathbf{z}_\mathbf{x}) + z_y \tag{27}$$
$$= \mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{z}_\mathbf{x} + z_y. \tag{28}$$

We know that when we sum two Gaussian random variables, its variance will also be summed. Therefore, we have

$$-\mathbf{w}^\top \mathbf{z}_\mathbf{x} + z_y \sim \mathcal{N}(0, 1 + \sum_i w_i^2) = \mathcal{N}(0, 1 + \|\mathbf{w}\|_2^2). \tag{29}$$

Therefore,

$$P_\mathbf{w}(y_i \mid \mathbf{x}_i) = \frac{1}{\sqrt{2\pi(\|\mathbf{w}\|_2^2 + 1)}} \exp\left(-\frac{1}{2}\frac{(y_i - \mathbf{w}^\top \mathbf{x})^2}{\|\mathbf{w}\|_2^2 + 1}\right). \tag{30}$$

The loss function can be computed by taking the log likelihood.

(e) **Implement the stochastic gradient descent to find the MLE for the model above.** In the starter code, we generate some data for you and you will need to recover $\mathbf{w}$ using the observed data. **Report the error** $\|w^* - w_{\text{true}}\|_2^2$ where w^* is the one that you recover. Try to play with hyper-parameters such as the batch size and the learning rate. **Is the solution of SGD sensitive to its hyper-parameters in this problem?**

Solution:

The error is `|w-w_true|**2 = 0.073309846`.

The SGD can easily stuck into a local minima with hyper-parameter settings. This indicates that it is likely that the problem is non-convex. This also tells us though SGD seems to be universal to all problems, it might not be a good choice for them. For problem such as TLS, SGD can solve it, but not as good compared the specialized algorithms such as SVD in term of the quality, reliability, and efficiency. Finally, I still need to emphasize that the objective function in the previous part is not as same as the objective function of SVD method, and therefore we cannot simply compare the solution.

4 [BONUS] Genome-Wide Association Study

All the following text is present in the accompanying jupyter notebook, but the notebook has additional explanations and accompanying figures and code. We recommend you do not read this pdf, but go directly to the jupyter notebook. This pdf was included for quick reference.

Overall goal: This real world problem is one in computational biology which uses many of the techniques and concepts you have been introduced to, all together, in particular, linear regression, PCA, non-iid noise, diagonalizing multivariate Gaussian covariance matrices, and bias-variance

trade-off. We will also tangentially introduce you to concepts of statistical testing. **This homework problem is effectively a demo in that we will ask you to execute code and answer questions about what you observe. You are not required to code anything at all.**

Setup and problem statement: Given a set of people for whom genetics (DNA) has been measured, and also a corresponding trait for each person, such as blood pressure, or "is-a-smoker", one can use data-driven methods to deduce which genetic effects are likely responsible for the trait. We have collected blood from n individuals who either smoke ($y_i = 1$) or do not smoke ($y_i = 0$). Their blood samples have been sequenced at m positions along the genome, yielding the feature matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, composed of the genetic variants (which take on values 0, 1, or 2). Specifically, $X_{i,j}$ is a numeric encoding of the DNA for the i^{th} person at genetic feature j. [1] We want to deduce which of the m genetic features are associated with smoking, by considering one at a time. *In the data we give you, it will turn out that there is no true signal in the data; however, we will see that without careful modelling, we would erroneously conclude that the data were full of signal.*

Overall modelling approach: The basic idea will be to "test" one genetic feature at a time and assign a score (a p-value) indicating our level of belief that it is associated with the trait. We will start with a simple linear regression model, and then build increasingly more correct linear predictive models. The three models we will use are (1) linear regression, (2) linear regression with PCA features, (3) linear regression with all genetic variants as features. The first model is naive, because it assumes the individuals are iid. The fundamental modelling challenges with these kinds of analyses is that the individuals in the study are not typically not iid, owing to differences in race and family-relatedness (e.g., sisters, brothers, patients, grandparents in the data set), which violate the iid assumption. Left unmodelled, such structure creates misleading results, yielding signal where none exists. Luckily, as we shall see, one can visualize these modelling issues via quantile-quantile plots, which will soon be briefly introduced.

How to test each variant: Herein we provide a minimal exposition to allow you to do the homework. To estimate how implicated each genetic feature is we will use a score, in the form of a p-value. One can think of the p-value as a proxy for how informative the genetic feature is for prediction of the trait (e.g. "is smoker"). More precisely, to get the p-value for the j^{th} genetic feature, we first include it in the model (for a given model class, such as linear regression) and compute the maximum likelihood, LL_j (this is our alternative hypothesis). Then we repeat this process, but having removed the genetic feature of interest from the model, yielding LL_{-j} (this is our null hypothesis). To be clear, the null hypothesis will have none of the m genetic variants that are being tested. You can refer to the jupyter notebook for a brief explanation of hypothesis testing. The p-value is then a simple monotonic decreasing function of the difference in these two likelihoods, $\texttt{diff\_ll} = LL_j - LL_{-j}$–one that we will give you. P-values lie in $[0, 1]$ and the smaller the p-value, the larger the difference in the likelihoods, and the more evidence that the genetic marker is associated with the trait (assuming the model is correct).

(a) To diagnose if something is amiss in our genetic analyses, we will make use of the following: (1) we assume that if any genetic signal is present, it is restricted to a small fraction of the

[1] Technically, the entries of the matrix correspond to having zero, one or two mutant versions of the DNA, but we will treat them as real-valued in this problem.

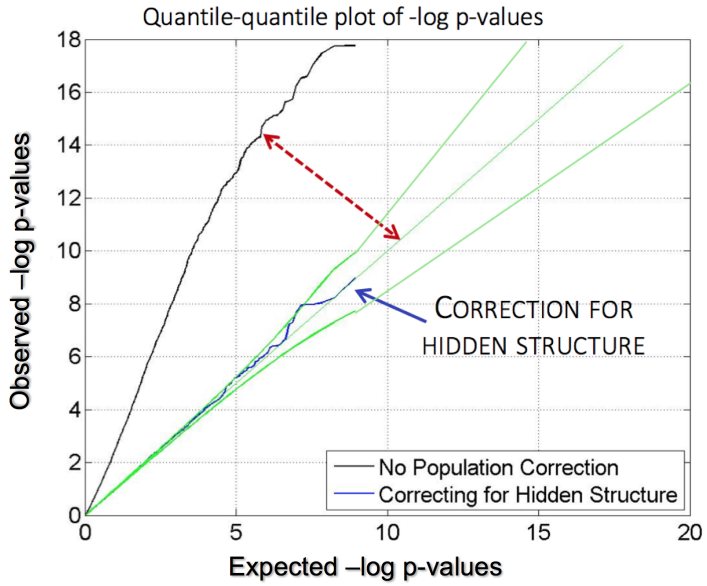


Figure 3: Example of quantile-quantile plot that shows large deviation from expectation.

genetic features, (2) p-values corresponding to no signal in the data are distributed as $p \sim$ Unif$[0, 1]$. Combining these two assumptions, we will assume that p-values arising from a valid model should largely be drawn from Unif$[0, 1]$, and also that large deviations suggest that our model is incorrect. Quantile-quantile plots let us visualize if we have deviations or not. Quantile-quantile plots are a way to compare two probability distributions by comparing their quantile values (e.g. how does the smallest p-value in each distribution compare, and then the second smallest, etc.). In the quantile-quantile plot, you will see m points, one for each genetic marker. The x-coordinate of each point corresponds to the theoretical quantile we would expect to see if the distribution was in fact a $Unif[0, 1]$ and the y-coordinate corresponds to the observed value of the quantile. An example is shown in Figure 1, where the line on the diagonal results from an analysis where the model is correct, and hence the theoretical and empirical p-value quantiles match, while the other line, which deviates from the diagonal, indicates that we have likely made a modelling error. If there are genetic signals in the data, these would simply emerge as a handful of outlier points from the diagonal (not shown).

Before we dive into developing our models, we need to be able to understand whether the p-values we get back from our hypothesis tests look like m random draws from a $Unif[0, 1]$.

Use the qqplot function to make a qq-plot for each of the 3 distributions provided below and explain your findings. What should we observe in our qq-plot if our empirical distribution looks more and more similar to a Unif[0, 1]?. Note that we use two kinds of qq-plots: one in p-value space and one negative log p-value space. The former is for intuition, while the latter is for higher resolution. The green lines in the negative log p-value qq-plots indicate the error bars.

Solution:

See Figure 5 and Figure 4 for plots on both the linear and negative log scale. As our empirical

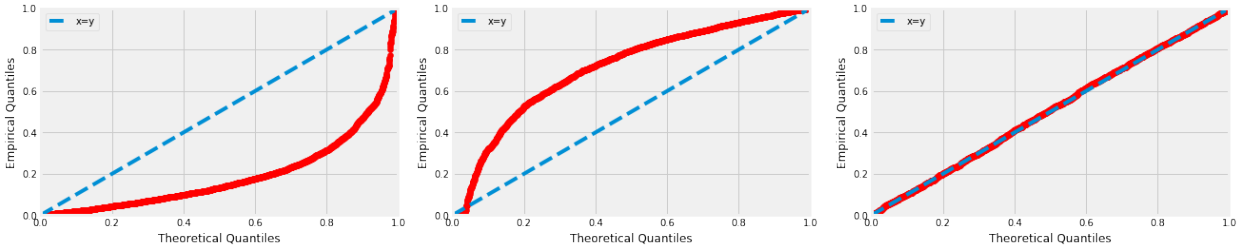


Figure 4: QQ-plot in linear scale of a skewed left, skewed right, and uniform distribution (from left to right).

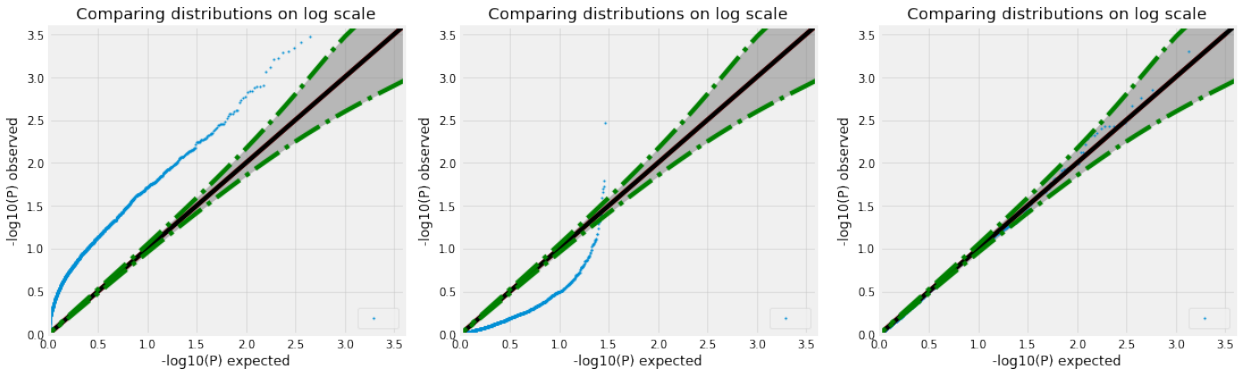


Figure 5: QQ-plot in -log scale of a skewed left, skewed right, and uniform distribution (from left to right).

distribution becomes more uniform, we should observe points that align along the diagonal in our qq-plot.

(b) We will use linear models in the genetic marker we are testing. In particular, when testing the j^{th} genetic feature, we have that the trait, $\mathbf{y}$, is a linear function of the genetic variant, i.e. $\mathbf{y} = \mathbf{x_j} w_1 + \mathbf{w_0} + \boldsymbol{\epsilon}$ where ϵ_i is random noise distributed as $N(0, \sigma^2)$, $w_1 \in \mathbb{R}^1$, $\mathbf{w_0} \in \mathbb{R}^{n \times 1}$ is a constant vector, and $\mathbf{x_j}$ is the jth column of $\mathbf{X}$, representing data for the jth genetic variant. To simplify matters, we will add a column of ones to the right end of $\mathbf{x_j}$ and rewrite the regression as $\mathbf{y} = [\mathbf{x_j}, \mathbf{1}]\mathbf{w} + \boldsymbol{\epsilon}$ where $[\mathbf{x_j}, \mathbf{1}]$ is the jth column of $\mathbf{X}$ with a vector of ones appended to the end and $\mathbf{w} \in \mathbb{R}^{2 \times 1}$. The model without any genetic information, $\mathbf{y} = \mathbf{1}w + \boldsymbol{\epsilon}$ is referred to as the *null model* in the parlance of statistical testing The *alternative model*, which includes the information we are testing (one genetic marker) is $\mathbf{y} = [\mathbf{x_j}, \mathbf{1}]\mathbf{w} + \boldsymbol{\epsilon}$ where $\mathbf{x_j}$ is the jth column of X, i.e. the data using only the jth genetic variant as a feature. **Plot the quantile-quantile plot of p-values using linear regression as just describe, a so-called naive approach, by running the function** `naive_model`. **From the plot, what do you conclude about the suitability of linear regression for this problem?**

Solution:

See Figure 6 shows m points, each corresponding to a genetic variantshows that all the points lie far outside the error bars. From this the model tells us that essentially every genetic marker is informative in predicting trait y, which is exactly at odds with what we might expect. This may be an artifact of correlations in our data not accounted for by the model.

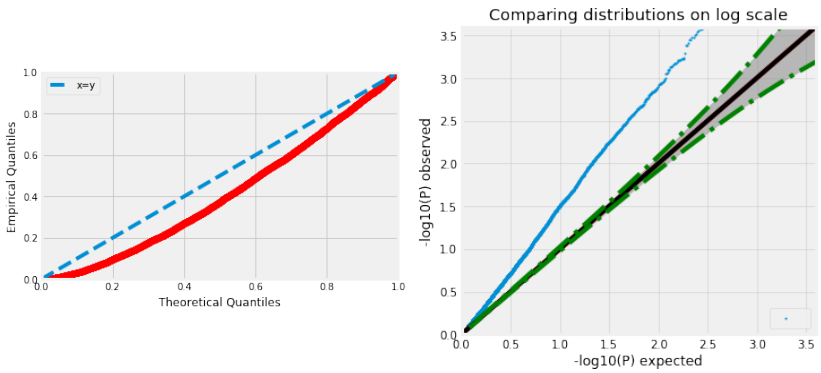


Figure 6: Figure of results from the naive model.

(c) From the quantile-quantile plot in the previous part, it appears that the model is picking up on more association than theoretically expected. The reason for this is owing to the assumption of iid noise. In particular, this data set contains individuals from different racial backgrounds, and also has clusters of individuals from extended families (e.g. grandparents, parents, siblings). This means that their genetics are not iid, and hence linear regression yields *spurious results*– all the genetic features seem to be implicated in the trait. Thus we need to modify our noise assumptions by somehow accounting for the hidden structure in the data set. The main idea is that when testing one genetic feature, all the other genetic features, jointly, are a proxy to the racial and family background. If we could include them in our model, we could correct the problem. **Ideally we would like to use all the genetic features in the linear regression model, however this is not a good idea. Why not?** Hint: There are roughly 1300 individuals and 7500 genetic variants. A written, English answer is sufficient.

So instead of using all genetic features, we will try using PCA to reduce the number of genetic features. As we saw in class, PCA on a genetic similarity matrix can capture geography, which correlates quite well to race. So instead of adding all the genetic features, we will instead use only three features[2], $\mathbf{X_{proj}}$, which are the $\mathbf{X}$ projected onto the top 3 principal components of $\mathbf{X}$. Consequently, the updated null model is $\mathbf{y} = \mathbf{X_{proj}}\mathbf{w_{proj}} + \boldsymbol{\epsilon}$ where $\mathbf{w_{proj}} \in \mathbb{R}^{3\times 1}$, while the alternative model is $\mathbf{y} = [\mathbf{x_j}, \mathbf{X_{proj}}, \mathbf{1}]\mathbf{w} + \boldsymbol{\epsilon}$ where $\mathbf{w} \in \mathbb{R}^{5\times 1}$ for genetic variant j. **Plot the quantile-quantile plot from obtaining p-values with this PCA linear regression approach by running the function** `pca_corrected_model`. **How does this plot compare to the first plot? What does this tell you about this model compared to the previous model?**

Solution: With the n, the number of individuals, being so much smaller than m, the number of genetic variants, this system is underdetermined. Even if it were not undetermined, we would need far, far more individuals than genetic markers for the bias-variance tradeoff not to hurt us. Using this model, $\mathbf{y} \sim N([\mathbf{x_j}, \mathbf{X_{proj}}, \mathbf{1}]\mathbf{w}, \sigma^2 I)$ for genetic variant j we get Figure 7.

This model clearly captures some of the confounding correlation and does a better job than the naive model, but the line still lies outside of the error bars.

(d) PCA got us part of the way there. However, PCA truncates the eigenspectrum; if the tail-end

[2]One needs to choose this number, but we have done so for you.

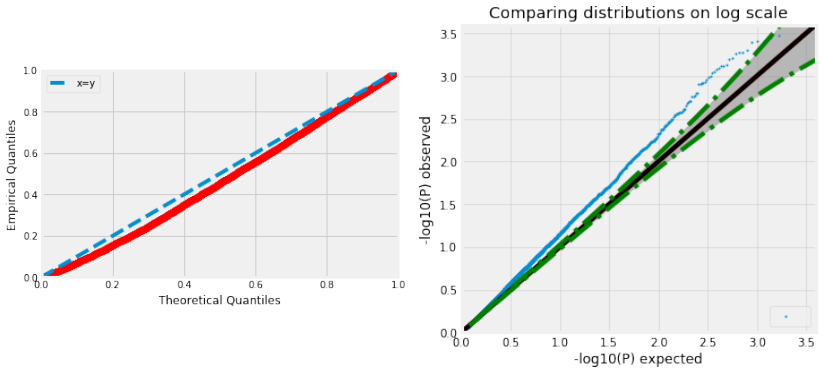


Figure 7: Figure of results from the model with features from PCA.

of that spectrum is important, as it is for family-relatedness, then it will not fully correct for our problem. So we want a method which (a) is well-behaved in terms of number of parameters that need to be estimated, and (b) includes all of the information we need. So rather than adding the projections as features, we use an modelling approach called linear mixed models which effectively adjust the iid noise in the gaussian by the pairwise genetic similarity of all the individuals. That is, we set Σ in $\mathbf{y} \sim N(y|[\mathbf{x_j}, 1]\mathbf{w}, I\sigma^2 + \mathbf{X}\mathbf{X}^\top\sigma_k^2)$.

Specifically, $\mathbf{y} = [\mathbf{x_j}, 1]\mathbf{w}+\mathbf{z}+\boldsymbol{\epsilon}$ where $\mathbf{z} \sim N(0, \sigma_k^2\mathbf{K})$ where $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$, σ_k, $\mathbf{w} \in \mathbb{R}^{m\times 1}$, and σ are parameters we want to estimate. Notice that $\mathbf{y} \sim N([\mathbf{x_j}, 1]\mathbf{w}, \sigma^2 I + \sigma_k^2 K)$. Evaluation of the likelihood is thus on the order of $O(n^3)$ from the required matrix inverse and determinant of $\sigma^2 I + \sigma_k^2 K$. To test m genetic variants, the time complexity becomes $O(mn^3)$, which is extremely slow for large datasets. **Given the eigen-decomposition $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, how can we make this faster if we have to test thousands of genetic feature? Would this be computationally more efficient if you only have one genetic feature to test?**

Finally, make the quantile-quantile plot for this last model by running the function `lmm`. What can we conclude about this model relative to the other two models?

HINT: Since the manipulations needed for $\sigma^2 I + \sigma_k^2 K$ is the bottleneck here, we would like a transformation which makes this covariance of the multi-variate gaussian be a diagonal matrix.

Solution:

$$\begin{aligned}
&\mathbf{y} \sim N([\mathbf{x_j}, 1]\mathbf{w}, \sigma_k^2\mathbf{K} + \sigma^2 I) \\
\implies &\mathbf{y} \sim N([\mathbf{x_j}, 1]\mathbf{w}, \sigma_k^2\mathbf{U}\mathbf{D}\mathbf{U}^\top + \sigma^2 I) \\
\implies &\mathbf{U}^\top\mathbf{y} \sim N(\mathbf{U}^\top[\mathbf{x_j}, 1]\mathbf{w}, \sigma_k^2\mathbf{D}^\top + \sigma^2 I) \\
\implies &\mathbf{U}^\top\mathbf{y} \sim N(\mathbf{U}^\top[\mathbf{x_j}, 1]\mathbf{w}, \sigma_k^2(\mathbf{D}^\top + \delta I))
\end{aligned}$$

where $\delta = \frac{\sigma^2}{\sigma_k^2}$. Alternatively,

$$\mathbf{y} = [\mathbf{x_j}, 1]\mathbf{w} + \mathbf{z} + \boldsymbol{\epsilon}$$

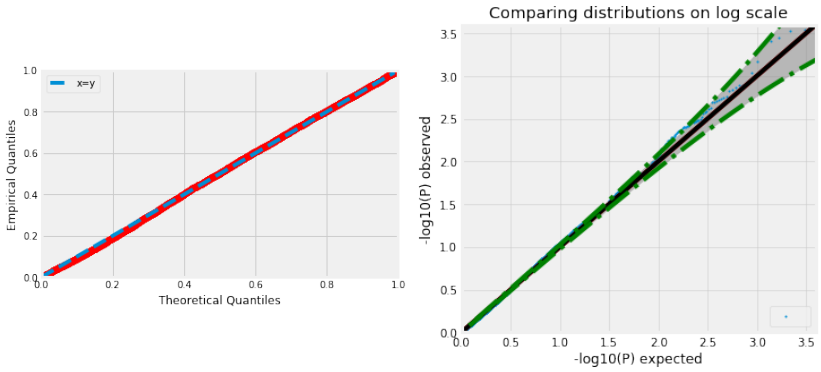


Figure 8: Figure of results from the model with features from LMM.

$$\mathbf{U}^\top \mathbf{y} = \mathbf{U}^\top [\mathbf{x_j}, 1]\mathbf{w} + \mathbf{U}^\top \mathbf{z} + \mathbf{U}^\top \boldsymbol{\epsilon}$$

which implies the mean is $E[\mathbf{U}^\top \mathbf{y}] = \mathbf{U}^\top \mathbf{X}\mathbf{w}$. For the variance,

$$\begin{aligned} Var(\mathbf{U}^\top \mathbf{y}) &= Var(\mathbf{U}^\top z) + Var(\mathbf{U}^\top \epsilon) \\ &= \mathbf{U}^\top (Var(z) + Var(\epsilon))\mathbf{U} \\ &= \mathbf{U}^\top (\sigma_k^2 \mathbf{K} + \mathbf{U}^\top \sigma^2 \mathbf{I})\mathbf{U} \\ &= \mathbf{U}^\top (\sigma_k^2 \mathbf{U}\mathbf{D}\mathbf{U}^\top + \mathbf{U}^\top \sigma^2 \mathbf{U}\mathbf{U}^\top)\mathbf{U} \\ &= \sigma_k^2 \mathbf{D} + \sigma^2 \mathbf{I} \end{aligned}$$

Now that we've diagonalized the covariance matrix, if we compute the SVD of K once, which is $O(n^3)$, and cache this result for future use, then we do not have to perform expensive inversions and determinant computations for each genetic variant we test. However, if we only test one marker, then this is the same amount of computational work. From Figure 8 this model does better than both our previous attempts. The line falls within our error bars and is aligned to what we expect to see given that the data has no true association signal.

5 Your Own Question

Write your own question, and provide a thorough solution.

Writing your own problems is a very important way to really learn the material. The famous "Bloom's Taxonomy" that lists the levels of learning is: Remember, Understand, Apply, Analyze, Evaluate, and Create. Using what you know to create is the top-level. We rarely ask you any HW questions about the lowest level of straight-up remembering, expecting you to be able to do that yourself. (e.g. make yourself flashcards) But we don't want the same to be true about the highest level.

As a practical matter, having some practice at trying to create problems helps you study for exams much better than simply counting on solving existing practice problems. This is because thinking

about how to create an interesting problem forces you to really look at the material from the perspective of those who are going to create the exams.

Besides, this is fun. If you want to make a boring problem, go ahead. That is your prerogative. But it is more fun to really engage with the material, discover something interesting, and then come up with a problem that walks others down a journey that lets them share your discovery. You don't have to achieve this every week. But unless you try every week, it probably won't happen ever.