601.481 Optimization: Machine Learning
Homework 1:
Jason Zhang
Collaborators: Archan Patel, Peter Klinkmueller, Will Ye
Due: Friday, October 26, 2018

1.1:

$$\begin{aligned} &\underset{\mathbf{U}\in\mathbb{R}^{dxk}}{minimize} : \mathbb{E}_{x\sim D}||x - \mathbf{U}\mathbf{U}^{\mathbf{T}}x||_2^2 \\ &\text{subject to: } \mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{I} \end{aligned}$$

To derive a stochastic gradient that is an unbiased estimator of the true gradient we can use the idea of a stochastic first order oracle in which we choose some g which is an unbiased estimator of the gradient at each iteration. Let $f(\mathbf{U}, x) = ||x - \mathbf{U}\mathbf{U}^{\mathbf{T}}x||_2^2$. Our choice for our stochastic gradient will be $\nabla f_n(\mathbf{U}, x)$. Since we are minimizing over $\mathbf{U}$, we will take the differential with respect to $\mathbf{U}$.

$$\begin{aligned} \frac{\partial f_n(\mathbf{U}, x)}{d\mathbf{U}} &= \frac{\partial((x - \mathbf{U}\mathbf{U}^{\mathbf{T}}x)^T(x - \mathbf{U}\mathbf{U}^{\mathbf{T}}x))}{d\mathbf{U}} \\ &= \frac{\partial(x^Tx - x^T\mathbf{U}\mathbf{U}^{\mathbf{T}}x)}{d\mathbf{U}} \\ &= -2xx^T\mathbf{U} \end{aligned}$$

Using linearity of expectations:

$$\begin{aligned} \mathbb{E}_{x\sim D}[\nabla f_n(\mathbf{U}, x)] &= \nabla\mathbb{E}_{x\sim D}[f_n(\mathbf{U}, x)] \\ &= \nabla\sum_{i=1}^{N} P(n = i)f_i(\mathbf{U}, x) \\ &= \nabla\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{U}, x) \\ &= \nabla f(\mathbf{U}, x) \end{aligned}$$

since our stochastic gradient satisfies the property of linearity of expectations, we know that our stochastic gradient is an unbiased estimator of the true gradient.

The Projected SGD algorithm is as follows:

1. Using our stochastic gradient, we iterate over our T input samples and compute $y_{i+1} = \mathbf{U}_i + \eta \nabla f(U_i, x)$
2. We then need to project our output from step 1 onto our constraint space which we do so by doing a QR decomposition of the output from step 1. QR decomposition uses Gram Schmidt orthogonalization to compute the decomposition and the Q matrix outputted will be an orthonormal matrix satisfying our constraint. We will set it as our x_{i+1} output.
3. The last iterate we get will be our optimum for our minimization problem.

1.2:
*Implemented in Jupyter notebook. The reported training error came out to about 3.78 (done by using the equation from the problem). I used a step size of 0.01 which gave me extremely fast convergence. I tried other options such as 1 and 0.001, but both led to much slower convergence at the same optimum as when i used a step size of 0.01.
1.3:
The reported test error was about 3.7841 (done by using the equation from the problem). The test set was created by doing a test-train split of the MNIST training data. Yes we can give a theoretical justification for the suboptimality of the returned iterate.
1.4:
Show the following problem is equivalent to the problem in 1.1:

$$\underset{\mathbf{P} \in \mathbb{R}^{dxd}}{minimize} : \mathbb{E}_{x \sim D} ||x - \mathbf{P}x||_2^2$$
$$\text{subject to: } \mathbf{P}^2 = \mathbf{P}, \mathbf{P}^T = \mathbf{P}, rank(\mathbf{P}) = k$$

We can start by setting $\mathbf{P} = \mathbf{U}\mathbf{U}^\mathbf{T}$. Now we have to prove that the constraints describe similar constraints in the first problem. Starting with the first constraint $\mathbf{P}^2 = \mathbf{P}$. Substituting $\mathbf{P}$ we get $(\mathbf{U}\mathbf{U}^\mathbf{T})^2 = (\mathbf{U}\mathbf{U}^\mathbf{T})\mathbf{U}\mathbf{U}^\mathbf{T} = \mathbf{U}\mathbf{U}^\mathbf{T}$. Thus this constraint holds in the first problem. We now move onto the second constraint: $\mathbf{P}^T = \mathbf{P}$. Substituting $\mathbf{P}$ we get $(\mathbf{U}\mathbf{U}^\mathbf{T})^T = (\mathbf{U}\mathbf{U}^\mathbf{T})$. Thus this constraint also holds. We move onto our final constraint in which $rank(\mathbf{P}) = k$. We start by conditioning on the dimensions of $\mathbf{U}$. We know that the dimensions are dxk, but given the constraint that $\mathbf{U}^\mathbf{T}\mathbf{U} = \mathbf{I}_{kxk}$ we know that $d >= k$ since the rank of a matrix is $min(d, k)$. With the fact that $\mathbf{U}^\mathbf{T}\mathbf{U} = \mathbf{I}_{kxk}$, it must be that $\mathbf{U}$ has a rank of k (since the Identity matrix has a full rank). Since the $rank(\mathbf{U}) = rank(\mathbf{U}^\mathbf{T}) = n$, we know by a property of rank that $rank(\mathbf{U}\mathbf{U}^\mathbf{T}) = rank(\mathbf{U}) = k$. Thus proving the last constraint. Therefore we have this is an equivalent problem to 1.1.
1.5:

$$\underset{\mathbf{P} \in \mathbb{R}^{dxd}}{minimize} : \mathbb{E}_{x \sim D} ||x - \mathbf{P}x||_2^2$$
$$\text{subject to: } \mathbf{P}^2 = \mathbf{P}, \mathbf{P}^T = \mathbf{P}, rank(\mathbf{P}) = k$$

In order for a problem to be considered a constrained convex optimization problem, each of

the constraints in the problem must also describe a convex set. In this case we can look at the constraint $\mathbf{P}^2 = \mathbf{P}$. This describes $\mathbf{P}$ being within the set of idempotent matrices. Let A $= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and let B $= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, we can take some convex combination of them such as: 0.5A + 0.5B $= \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$. This is not an idempotent matrix and thus $\mathbf{P}^2 = \mathbf{P}$ is not a convex set constraint, and therefore our problem does not fit in the convex optimization framework.

1.6:

Show 1.4 is equivalent to:

$$\begin{aligned} &\underset{\mathbf{P} \in \mathbb{R}^{dxd}}{maximize} : \mathbb{E}_{x \sim D}[\langle \mathbf{P}, xx^T \rangle] \\ &\text{subject to: } \mathbf{P}^2 = \mathbf{P}, \mathbf{P}^T = \mathbf{P}, rank(\mathbf{P}) = k \end{aligned}$$

$\langle \mathbf{P}, xx^T \rangle = ||x - \mathbf{P}x||_2^2$. To prove these are the same, we need to prove that the objectives for optimization are the same. Starting with $\langle \mathbf{P}, xx^T \rangle$, we can rewrite it as $\langle \mathbf{P}, xx^T \rangle = trace(\mathbf{P}^T xx^T) = trace(\mathbf{P}xx^T) = x^T\mathbf{P}^T x = x^T\mathbf{P}x$(since $\mathbf{P}x$ is a vector we can use the vectorize property of trace). We can now start by proving that the minimization problem optimizes the same objective. Since $\mathbf{P}$ is an idempotent and symmetric matrix, we have that it is a projection matrix. Given that we know that $||x||^2 = ||x - \mathbf{P}x||^2 + ||\mathbf{P}x||^2$ (an x vector projected with P), rearranging we get: $||x - \mathbf{P}x||^2 = ||x||^2 - ||\mathbf{P}x||^2$. Our problem in 1.4 minimizes $||x - \mathbf{P}x||^2$ and thus within our equation, minimizing $||x - \mathbf{P}x||^2$ means that we should maximize the difference by maximizing $||\mathbf{P}x||^2$. Thus our problem now can be shifted to trying to maximize $||\mathbf{P}x||^2 = (\mathbf{P}x)^T(\mathbf{P}x) = x^T\mathbf{P}^T\mathbf{P}x = x^T\mathbf{P}\mathbf{P}x = x^T\mathbf{P}x$. Both problems are trying to maximize $x^T\mathbf{P}x$ under the same constraints which means they are equivalent problems.

1.7:

Show 1.6 is equivalent to:

$$\begin{aligned} &\underset{\mathbf{P} \in \mathbb{R}^{dxd}}{maximize} : \mathbb{E}_{x \sim D}[\langle \mathbf{P}, xx^T \rangle] \\ &\text{subject to: } 0 \preceq \mathbf{P} \preceq \mathbf{I}, \mathbf{P}^2 = \mathbf{P}, \mathbf{P}^T = \mathbf{P}, Trace(\mathbf{P}) = k \end{aligned}$$

We need to show that the set A $\{\mathbf{P} \in \mathbb{R}^{dxd} : Trace(\mathbf{P}) = k, \mathbf{P}^T = P, 0 \preceq \mathbf{P} \preceq I\}$ is convex and its extreme points are exactly the set $\{P \in \mathbb{R}^{dxd} : \mathbf{P}^2 = P, \mathbf{P}^T = \mathbf{P}, rank(\mathbf{P}) = k\}$. Starting with convexity, we choose two matrices that satisfy the constraint set which we will denote as U, V. We will prove that any convex combination of U, V will still be closed under each of the constraints. Starting with $\mathbf{P}^T = P, 0 \preceq \mathbf{P} \preceq I$, this gives us a positive semi definite ordering constraint on our matrices. We choose some $\lambda \in (0, 1)$ and construct a convex combination with U, V which will give us $M = \lambda U + (1 - \lambda)V$. With our constraint we know that $0 \leq z^T Uz \leq 1$ and $0 \leq z^T Vz \leq 1 \forall z \in \mathbb{R}$. Using λ we can make a convex combination like such: $0 \leq \lambda z^T Uz + (1-\lambda)z^T Vz \leq 1 - \lambda + \lambda = 1$. We can simplify it to $0 \leq z^T(\lambda U + (1-\lambda)V)z \leq 1$, which is: $0 \leq z^T(M)z \leq 1$, thus $0 \preceq M \preceq \mathbf{I}$ and the positive semi definite order constraint holds. Proving the transpose constraint: $(\lambda U + (1-\lambda)V)^T = \lambda U^T + (1-\lambda)V^T = \lambda U + (1-\lambda)V$, and thus it still holds. Proving the trace constraint: Given that $tr(V)$ and $tr(U)$ are both equal to k, we have that $tr(\lambda U + (1-\lambda)V) = tr(\lambda U) + tr((1-\lambda)V) = \lambda k + (1-\lambda)k = k$, thus the trace constraint holds. Since all three constraints hold we have that our constraints form a

convex set.
We now want to show the extreme points are the set B $\{\mathbf{P} \in \mathbb{R}^{dxd} : \mathbf{P}^2 = \mathbf{P}, \mathbf{P}^T = \mathbf{P}, rank(\mathbf{P}) = k\}$. Assuming that we do not have an extreme point set that means that B will contain points from set A making $B \subset A$. Let us choose two matrices from set A that satisfy the conditions of set A which we will denote as C and D. Since $B \subset A$ we should be able to make a convex combination using C and D that will give us a point in set B. However the set of idempotent matrices does not form a convex set. Therefore we cannot have the set B as a subset of set A since without the bound of convexity, B can potentially include a matrix not within the set, thus the set B is unreachable from convex combinations of points in set A.

1.8:
To derive a stochastic gradient that is an unbiased estimator of the true gradient for problem 1.7 we can use the idea of a stochastic first order oracle in which we choose some g which is an unbiased estimator of the gradient at each iteration. Let $f(\mathbf{P}, x) = \langle \mathbf{P}, xx^T \rangle$. Our choice for our stochastic gradient will be $\nabla f_n(\mathbf{P}, x)$. Since we are maximizing over $\mathbf{P}$, we will take the differential with respect to $\mathbf{P}$.

$$\begin{aligned}
\frac{\partial f_n(\mathbf{P}, x)}{d\mathbf{P}} &= \frac{\partial(\langle \mathbf{P}, xx^T \rangle)}{d\mathbf{P}} \\
&= \frac{\partial(trace(\mathbf{P}xx^T))}{d\mathbf{P}} \\
&= (xx^T)^T \\
&= xx^T
\end{aligned}$$

Using linearity of expectations:

$$\begin{aligned}
\mathbb{E}_{x\sim D}[\nabla f_n(\mathbf{P}, x)] &= \nabla \mathbb{E}_{x\sim D}[f_n(\mathbf{P}, x)] \\
&= \nabla \sum_{i=1}^{N} P(n = i) f_i(\mathbf{P}, x) \\
&= \nabla \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{P}, x) \\
&= \nabla f(\mathbf{P}, x)
\end{aligned}$$

since our stochastic gradient satisfies the property of linearity of expectations, we know that our stochastic gradient is an unbiased estimator of the true gradient.

The version of SGD we will give is similar to 1.1 which is:

1. Using our stochastic gradient, we iterate over our T input samples and compute $y_{i+1} = \mathbf{P}_i + \eta \nabla f(\mathbf{P}_i, x)$
2. We then need to project our output from step 1 onto our constraint space which we do so by doing a QR decomposition of the output from step 1. QR decomposition uses Gram Schmidt orthogonalization to compute the decomposition and the Q matrix outputted

will be an orthonormal matrix. This will satisfy our constraints of having eigenvalues of either 0 or 1, the constraint that the matrix is symmetric(which will be satisfied due to our matrix being positive semi definite), and the rank will be k. We will set it as our x_{i+1} output.

3. The last iterate we get will be our optimum for our minimization problem.

1.9:
Prove algorithm returns an $\epsilon-$suboptimal solution in time $O(1/\epsilon^2)$.

Proof:
We want to be able to bound $E[f(\hat{\mathbf{P}})] - f(\mathbf{P}^*) \leq \epsilon$, where $f(\mathbf{P}) = \langle \mathbf{P}, xx^T \rangle$. We can start from Jensen's inequality,

$$f(\hat{\mathbf{P}}) - f(\mathbf{P}^*) \leq \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{P}_t) - f(\mathbf{P}^*)$$

Now we aim to bound $\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{P}_t) - f(\mathbf{P}^*)$. To do so, we start by trying to bound the suboptimality of the Ith iterate. We will let $Y_{t+1} = \mathbf{P}_t - \eta g_t$. η is our step size, g_t is our gradient for some sample, and $\mathbf{P}_t$ is our matrix at some t. We can assume that our data is scaled in such a way that $||x||^2 \leq C$ for some constant C. When computing $||g_t||^2$ we know that $g_t = xx^T$ and thus $||g_t||^2 = ||x||^4$ which is also constant bound by C^2.

$$\begin{aligned}
f(\mathbf{P}_t) - f(\mathbf{P}) &\leq \langle g_t, P_t - P^* \rangle && \text{by convexity} \\
&= \frac{1}{\eta}\langle \mathbf{P}_t - Y_{t+1}, \mathbf{P}_t - P* \rangle && \text{using } Y_{t+1} = \mathbf{P}_t - \eta g_t \\
&= \frac{1}{2\eta}(||\mathbf{P}_t - Y_{t+1}||^2 + ||\mathbf{P}_t - \mathbf{P}^*||^2 - ||\mathbf{P}^* - Y_{t+1}||^2) && \text{using } u^T v = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u - v\|^2) \\
&= \frac{\eta}{2}||g_t||^2 + \frac{1}{2\eta}(||\mathbf{P}_t - \mathbf{P}^*||^2 - ||\mathbf{P}^* - Y_{t+1}||^2) && \text{using } Y_{t+1} = \mathbf{P}_t - \eta g_t \\
&\leq \frac{\eta}{2}||g_t||^2 + \frac{1}{2\eta}(||\mathbf{P}_t - \mathbf{P}^*||^2 - ||\mathbf{P}_{t+1} - \mathbf{P}^*||^2) && \text{projection property} \\
&\leq \frac{\eta}{2}C^2 + \frac{1}{2\eta}(||\mathbf{P}_t - \mathbf{P}^*||^2 - ||\mathbf{P}_{t+1} - \mathbf{P}^*||^2) && ||g_t||^2 \leq C^2 \text{from earlier}
\end{aligned}$$

Now we use the above property to show for across all T:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T} f(\mathbf{P}_t) - f(\mathbf{P}^*) &\leq \frac{1}{T}\sum_{t=0}^{T}\frac{C^2\eta}{2} + \frac{1}{T}\sum_{t=0}^{2\eta T} ||\mathbf{P}_t - \mathbf{P}^*||^2 - ||\mathbf{P}_{t+1} - \mathbf{P}^*||^2 \\
&= \frac{C^2\eta}{2} + \frac{1}{2\eta T}||\mathbf{P}_t - \mathbf{P}^*||^2 - ||\mathbf{P}_{t+1} - \mathbf{P}^*||^2 \\
&\leq \frac{C^2\eta}{2} + \frac{1}{2\eta T}||\mathbf{P}_t - \mathbf{P}^*||^2 \\
&\text{using } (||\mathbf{P}_t - \mathbf{P}^*|| \geq 0) \\
&\leq \frac{C^2\eta}{2} + \frac{1}{2\eta T}||\mathbf{P}^*||^2 \\
&= \frac{C^2\eta}{2} + \frac{1}{2\eta T}k \\
&\text{(using fact that P contains k eigenvalues bounded to be 0 or 1, thus)} ||P^*||^2 = k
\end{aligned}$$

Now that we have a bound, we can bound our first equation in the beginning by it which gives us:

$$\begin{aligned}
f(\hat{\mathbf{P}}) - f(\mathbf{P}^*) &\leq \frac{C^2\eta}{2} + \frac{1}{2\eta T}k \\
&\leq \frac{C^2 k + 1}{2\sqrt{T}} \qquad \text{for } \eta = k/\sqrt{T}
\end{aligned}$$

Therefore to achieve our ϵ suboptimality, T needs to be some factor of ϵ^2, which means our algorithm returns an ϵ suboptimal solution in $O(1/\epsilon^2)$ time.

1.10:

It is asymptotically optimal as shown in class.

2.1:

Show that the objective in problem A.4 is λ-strongly convex w.r.t. the l_2 norm.

We first let $h(x) = f(x) - \frac{\lambda}{2}||x||_2^2$ and show that it is convex as by Proposition 3.2.21. From our problem we have that $f(w) = \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{m}\sum_{(x,y)\in S} l(w;(x,y)) \quad , \quad l(w;(x,y)) = max\{0, 1 - y\langle w, x\rangle\}$. Thus our $h(w) = \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{m}\sum_{(x,y)\in S} l(w;(x,y)) - \frac{\lambda}{2}\|x\|_2^2 = \frac{1}{m}\sum_{(x,y)\in S} l(w;(x,y)) = \frac{1}{m}\sum_{(x,y)\in S} max\{0, 1 - y\langle w, x\rangle\}$.

To test convexity we need to show that:

$$\forall x_1, x_2 \in X, \forall \lambda \in [0,1] : f(\lambda x_1 + (1-\lambda)x_2 \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

So we test by using some t between 0 and 1 and some vectors $u, v \in \mathbb{R}^n$.

$$\begin{aligned}
h(tu+(1-t)v) &= \frac{1}{m}\sum_{(x,y)\in S} max\{0, 1-y\langle tu+(1-t)v, x\rangle\} \\
&= \frac{1}{m}\sum_{(x,y)\in S} max\{0, 1-ty\langle u, x\rangle-(1-t)y\langle v, x\rangle\} \\
&= \frac{1}{m}\sum_{(x,y)\in S} max\{0, 1-ty\langle u, x\rangle-(1-t)y\langle v, x\rangle+t-t\} \\
&= \frac{1}{m}\sum_{(x,y)\in S} max\{0, t-ty\langle u, x\rangle(1-t)-(1-\lambda)y\langle v, x\rangle\} \\
&= \frac{1}{m}\sum_{(x,y)\in S} max\{0, t-ty\langle u, x\rangle\}+\frac{1}{m}\sum_{(x,y)\in S} max\{0, (1-t)-(1-t)y\langle v, x\rangle\} \\
&= t\frac{1}{m}\sum_{(x,y)\in S} max\{0, 1-y\langle u, x\rangle\}+(1-t)\frac{1}{m}\sum_{(x,y)\in S} max\{0, 1-(1-y\langle v, x\rangle\} \\
&= tf(u)+(1-t)f(v)
\end{aligned}$$

Thus we have satisfied the inequality in which $h(\lambda u+(1-\lambda)v) \leq \lambda f(u)+(1-\lambda)f(v)$, and therefore $h(w)$ is convex. Since $h(w)$ is convex then $f(w)$ is $\lambda-$strongly convex, by the proposition.

2.2:

Compute the sub-gradient for the objective in problem A.4

Since $l(w;(x,y)) = max\{0, 1-y\langle w, x\rangle\}$ can take on two different values, we will condition on if $1-y\langle w, x\rangle \geq 0$ or $1-y\langle w, x\rangle < 0$.

If $1-y\langle w, x\rangle < 0$, then we have that $l(w;(x,y)) = 0$. Thus we get:

$$\begin{aligned}
f(w) &= \frac{\lambda}{2}\|w\|_2^2+\frac{1}{m}\sum_{(x,y)\in S} 0 \\
&= \frac{\lambda}{2}\|w\|_2^2
\end{aligned}$$

Now we take the subgradient of $f(w)$:

$$\frac{d(f(w))}{\partial w} = \lambda w$$

If $1-y\langle w, x\rangle \geq 0$, then we have that $l(w;(x,y)) = y\langle w, x\rangle$. Thus we get:

$$f(w) = \frac{\lambda}{2}\|w\|_2^2+\frac{1}{m}\sum_{(x,y)\in S} 1-y\langle w, x\rangle$$

As we are dealing with subgradients we consider a gradient for every sample separately and

thus our equation to differentiate becomes: $f_t(w) = \frac{\lambda}{2}\|w\|_2^2 + 1 - y_t\langle w, x\rangle$. We compute the gradient for f_t:

$$\begin{aligned}\frac{d(f(w))}{\partial w} &= \frac{d(\frac{\lambda}{2}\|w\|_2^2 + 1 - y_t w^T x_t)}{\partial w} \\ &= \lambda w - y_t^T x_t\end{aligned}$$

Our subgradient is: $\begin{cases}\lambda w - y_t^T x_t & (1 - y\langle w, x\rangle) \geq 0 \text{ for some sample t} \\ \lambda w & (1 - y\langle w, x\rangle) < 0\end{cases}$

2.3:

We begin by following the proof for SGD and bound the iterates. We will have w be our result at some iteration and u be our optimal solution.
Let $\Delta_t = \mathbb{E}_{g_t}[g_t|w_t]$. Since $\Delta_t \in \partial f_i(w_t)$ and f_i is strongly convex, then we can say:

$$f_i(w_t) - f_i(u) \leq \langle \Delta_t, w_t - u\rangle - \frac{\lambda}{2}\|w_t - u\|^2$$

We can bound $\langle \Delta_t, w_t - u\rangle$ as follows:

$$\begin{aligned}\langle g_t, w_t - u\rangle &= \frac{1}{\eta_t}(w_t - y_{t+1})^T(w_t - u) && \text{by the update rule} \\ &= \frac{1}{2\eta_t}(\|w_t - u\|2 + \|w_t - y_{t+1}\|^2 - \|y_{t+1} - u\|^2) && u^T v = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u - v\|^2) \\ &= \frac{1}{2\eta_t}(\|w_t - u\|2 - \|y_{t+1} - u\|^2) + \frac{\eta_t}{2}\|g_T\|^2 && \text{by the update rule} \\ &\leq \frac{1}{2\eta_t}(\|w_t - u\|2 - \|w_{t+1} - u\|^2) + \frac{\eta_t}{2}\|g_T\|^2 && \text{projection property}\end{aligned}$$

We know that since our functions are G-lipschitz then $\mathbb{E}[||g_t||^2] \leq G^2$ Now taking the expectation of both sides with respect to g_t:

$$\langle \Delta_t, w_T - u\rangle \leq \mathbb{E}[\frac{\|w_t - u\|2 - \|w_{t+1} - u\|^2}{2\eta_t}] + \frac{\eta_t G^2}{2}$$

We can now plug this bound back into the first equation and sum over $t = 1...T$ to get:

$$\sum_{t=1}^{T}\mathbb{E}[f_t(w_t)] - f_t(u) \leq \mathbb{E}[\sum_{t=1}^{T}\frac{\|w_t - u\|2 - \|w_{t+1} - u\|^2}{2\eta_t} - \frac{\lambda}{2}\|w_t - u\|^2] + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t$$

Substituting $\eta_t = (1/\lambda t)$:

$$\sum_{t=1}^{T}\mathbb{E}[f_t(w_t)] - f_t(u) \leq \mathbb{E}[\sum_{t=1}^{T}\frac{\lambda}{2}((t-1)\|w_t - u\|^2 - t\|w_{t+1} - u\|^2)] + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t$$

The term that survives in the first term on the right hand side is $-\frac{\lambda T}{2}\|w_T - u\|^2 \leq 0$
Therefore we have:

$$\sum_{t=1}^{T} \mathbb{E}[f_t(w_t)] - f_t(u) \leq \frac{G^2}{2\lambda} \sum_{t=1}^{T} \frac{1}{t} \leq \frac{G^2}{2\lambda}(1 + log(T))$$

This is equivalent to:

$$\sum_{t=1}^{T} f_t(w_t) \leq \sum_{t=1}^{T} f_t(u) + \frac{G^2}{2\lambda}(1 + log(T))$$

Dividing by T we get our final answer:

$$\frac{1}{T}\sum_{t=1}^{T} f_t(w_t) \leq \frac{1}{T}\sum_{t=1}^{T} f_t(u) + \frac{G^2}{2\lambda T}(1 + log(T))$$

2.4:
Assume that $\|x\| \leq R \quad \forall x \in S$. Show that the subgradients of the objective function in (A.4) are bounded in norm. Give the bound.
We can start with that our w at every iteration is bounded by $B_0(\frac{1}{\sqrt{\lambda}})$. This means that at every iteration $||w|| \leq \frac{1}{\sqrt{(\lambda)}}$.
Since $l(w;(x,y)) = max\{0, 1 - y\langle w, x\rangle\}$ can take on two different values, we will condition on if $1 - y\langle w, x\rangle \geq 0$ or $1 - y\langle w, x\rangle < 0$.
Our subgradient is: $\begin{cases} \lambda w - y_t^T x_t & (1 - y\langle w, x\rangle) \geq 0 \text{ for some sample t} \\ \lambda w & (1 - y\langle w, x\rangle) < 0 \end{cases}$
For the subgradient: λw, we know that $w \leq ||w||$, and thus we get:

$$\begin{aligned} \lambda w &\leq \lambda||w|| \\ &\leq \lambda\frac{1}{\lambda} && \text{defined earlier} \\ &= \sqrt{\lambda} \end{aligned}$$

For the subgradient: $\lambda w - y_t^T x_t$, we know that it is bounded by its norm which gives us:

$$\begin{aligned} \lambda w - y_t^T x_t &\leq ||\lambda w - y_t^T x_t|| \\ &\leq ||\lambda w + x|| \\ &\leq ||\lambda w|| + ||x|| \\ &\leq \sqrt{\lambda} + R && \text{from earlier calculations and given} \end{aligned}$$

So our bound for the subgradient λw is $\sqrt{\lambda}$, and our bound for the subgradient $\lambda w - y_t^T x_t$ is $\sqrt{\lambda} + R$.

2.5:

Assume that the optimum w^* of Problem A.4 is in $B_0 \frac{1}{\sqrt{\lambda}}$. Derive a bound on sub-optimality of the output $\bar{w}$.

We want to be able to bound $\mathbb{E}[f(w)] - f(u) \leq \epsilon$. From our work in 2.3 we concluded that $\frac{1}{T}\sum_{t=1}^{T} f_t(w_t) - \frac{1}{T}\sum_{t=1}^{T} f_t(u) \leq \frac{G^2}{2\lambda T}(1 + log(T))$. Thus in order to guarantee some sort of ϵ sub-optimiality T will have to be some factor of $1/\epsilon$, thus our algorithm will have a complexity bound of $O(1/\epsilon)$.