Name: Hongda Li
AMATH 516 HW4 2021 FALL

4.2

This is the Fenchel Rockafella Primal Dual Problems:

$$
\begin{aligned}
(P)\ &\inf_{x\in\mathbf{E}}\left\{h(\mathcal{A}x)+g(x)\right\} \\
(D)\ &\sup_{y\in\mathbf{Y}}\left\{-h^\star - g^\star(-\mathcal{A}^*y)\right\}
\end{aligned} \tag{4.2.1}
$$

Where we use $\star$ for the Fenchel Conjugate of the function and $*$ for adjoint operator.
Sometimes we need to take the subdifferential/gradient/conjugate of an expression involving multiple parameters, we denote it as the following:

$$\nabla[f(x,y)|x] \quad \partial[f(x,y)|x] \quad [f(x,y)|x]^\star$$

Where the bar notation is telling use which variable we are taking the derivative wrt to the parameter.

4.2.1

We wish to compute the Fenchel Rockafella's Dual for the following Optimization Problem:

$$\min_x\left\{\frac{1}{2}\left\|Ax-b\right\|_2^2+\|x\|_1\right\}$$

To compute the Dual first identify f, g:

$$
\begin{aligned}
h(Ax) &= \frac{1}{2}\|Ax-b\|_2^2 \\
\implies h(x) &= \frac{1}{2}\|x-b\|_2^2 \\
g(x) &= \|x\|_1
\end{aligned} \tag{4.2.1.1}
$$

Considering the Fenchel Conjugate of $h(x)$:

$$
\begin{aligned}
h^\star(x) &= \sup_y\left\{\langle x,y\rangle - h(y)\right\} \\
&= \sup_y\left\{\langle x,y\rangle - \frac{1}{2}\left\|y-b\right\|_2^2\right\} \\
\partial\left[\langle x,y\rangle - \frac{1}{2}\|y-b\|_2^2\middle| y\right] &= \{x-(y-b)\} \\
0 &\in \{x-(y^+-b)\} \\
\implies y^+ &= x+b \\
\implies h^\star(x) &= \langle x, x+b\rangle - \frac{1}{2}\left\|x\right\|_2^2 \\
&= \langle x,x\rangle + \langle x,b\rangle - \frac{1}{2}\left\|x\right\|_2^2 \\
&= \|x\|_2^2 + \langle x,b\rangle - \frac{1}{2}\left\|x\right\|_2^2 \\
h^\star(x) &= \frac{1}{2}\|x\|_2^2 + \langle x,b\rangle
\end{aligned} \tag{4.2.1.2}
$$

In this case, the function is continuous and it's concave, therefore a minimum eixsts and if I take the derivative and set it zero, I can get the optimal paramter y^+ and find the supremum.
Next we compute the Conjugate for $g(x)$:

$$\begin{aligned} g(x) &= \|x\|_1 \\ g^\star(x) &= \sup_{y} \langle \langle x, y\rangle - \|y\|_1 \rangle \\ &= \sup_{y} \left\{ \sum_{i=1}^{n} (x_i y_i - |y_i|) \right\} \\ &= \sum_{i=1}^{n} \sup_{y_i} \{x_i y_i - |y_i|\} \end{aligned} \tag{4.2.1.3}$$

The sup can goes in because it's a summing up discrate expression each using y_i, and we assume that $y \in \mathbb{R}^n$, and we can say that:

$$\begin{aligned} & \forall 1 \le i \le n \\ & \sup_{y_i} \{x_i y_i - |y_i|\} = \max \left\{ \sup_{y_i \ge 0} \{x_i y_i - y_i\}, \sup_{y_i < 0} \{x_i y_i + y_i\} \right\} \\ & \sup_{y_i > 0} \{y_i(x_i - 1)\} = \begin{cases} \infty & x_i > 1 \\ 0 & x_i \le 1 \end{cases} \\ & \sup_{y_i < 0} \{y_i(x_i + 1)\} = \begin{cases} 0 & x_i \ge -1 \\ \infty & x_i < -1 \end{cases} \\ \implies \forall 1 \le i \le n \quad & \sup_{y_i} \{x_i y_i - |y_i|\} = \begin{cases} 0 & |x_i| \le 1 \\ \infty & |x_i| > 1 \end{cases} \\ \implies \sum_{i=1}^{n} & \sup_{y_i} \{x_i y_i - |y_i|\} = \begin{cases} \infty & \|x\|_\infty > 1 \\ 0 & \|x\|_\infty \le 1 \end{cases} \\ & \implies g^\star(x) = \delta\{\|x\|_\infty \le 1\}(x) \end{aligned} \tag{4.2.1.4}$$

The conjugate of g is the indicator function of the unit infinity norm ball.
Now we are ready to use the conjugate function to figure out the Fenchel Rockafella's Dual, we use the results from (4.2.1.2) and (4.2.1.4):

$$\begin{aligned} & \sup_{y} \{-h^\star(y) - g^\star(-\mathcal{A}^* y)\} \\ \equiv & \sup_{y} \left\{ -\frac{1}{2}\|y\|_2^2 - \langle y, b\rangle - \delta\{\| - A^T y\|_\infty \le 1\}(x) \right\} \\ \equiv & \sup_{\|A^T y\|_\infty \le 1} \left\{ -\frac{1}{2}\|x\|_2^2 - \langle x, b\rangle \right\} \end{aligned} \tag{4.2.1.5}$$

4.2.2

We wish to look for the Dual for the following problem:

$$\min_{\|x\|_q \le 1} \|Ax - b\|_p$$

Where we are going to use $\|x\|_m, \|x\|_{\bar{m}}$ to denote the norm and its dual norm. First we identify the f, g as:

$$\begin{aligned} h(x) &= \|x - b\|_p \\ g(x) &= \delta\{\|x\|_q \le 1\}(x) \end{aligned} \tag{4.2.2.1}$$

Before we start I am going to introduce a convenient new trick and I am going to prove it:
Claim: “Conjugate of a Norm is the indicator function for the unit norm ball of the Dual Norm.”
Proof:
Set $f(x) = \|x\|$ and we use $\|x\|_*$ to denote the dual norm, and we wish to prove that $f^\star(x) = \delta\{\|x\|_* \le 1\}(x)$. Recall that the definition of dual norm is:

$$\|y\|_* = \sup_{\|x\|\le 1} \langle x, y\rangle$$

From the definition of the conjugate of a function we have: $f^\star(x) = \sup_y\{\langle x, y\rangle - f(y)\}$, and let's discuss it by cases:
Suppose that $\|y\|_* \le 1$ then:

$$\begin{aligned}
1 \ge \|y\|_* \ge \sup_{\|x\|\le 1} \{\langle x, y\rangle\} \\
\forall \|x\| \le 1 : \langle x, y\rangle \le 1 \\
\text{let: } t > 0 \\
\langle tx, y\rangle \le t\langle x, y\rangle \le t = t\|x\| = \|tx\| \\
\Longrightarrow \ \forall\ tx\ , \|x\| \le 1 : \langle tx, y\rangle \le \|tx\| \\
\Longrightarrow \ \forall x\ \langle x, y\rangle \le \|x\| \\
\Longrightarrow \ \sup_x \{\langle y, x\rangle - \|x\|\} = 0
\end{aligned} \tag{4.2.2.2}$$

Because the inner produce is forever smaller than the norm of x, therefore we have to set the value of $x = 0$ to attain the sup of the expression, which is zero.

Next we consider the case where $\|y\|_* > 1$, and we can say that:

$$\begin{aligned}
\|y\|_* > 1 \\
\exists x :\ \|x\| \le 1 \wedge \langle x, y\rangle > 1 \\
\Longrightarrow \ \sup_y \{\langle x, y\rangle - \|x\|\} \ge \langle x, y\rangle - \|x\| > 0 \\
\text{let } x = tx, t \ge 0 \\
t\langle x, y\rangle - t\|x\| > 0 \\
t(\langle x, y\rangle - \|x\|) > 0 \\
\Longrightarrow \ \sup_y \{\langle x, y\rangle - \|x\|\} = \infty
\end{aligned} \tag{4.2.2.3}$$

Therefore the conjugate of the function f is: $f^\star(x) = \delta\{\|x\|_* \le 1\}(x)$. It's the indicator function of the dual unit norm ball.

Using this pieces of information, we can start figuring out the conjugate. From table 3.4 and the fact that the function f, h are both norm of something, hence convex, therefore the double conjugate of the function is the funciton itself. Therefore the table can be applied in both forward and backwards direction. In our case, we are applying the seceond entry of table 3.4.

$$[f(x+b)|\, x]^\star = f^\star(y) - \langle b, y\rangle \tag{4.2.2.4}$$

In our case we have:

$$[\,\|x-b\|_p|\, x]^\star\,(y) = \delta\{\|x\|_{\bar{p}} \le 1\}(y) - \langle b, y\rangle \tag{4.2.2.5}$$

By a similar token the conjugate of the g function would be:

$$[\delta\{\|x\|_{q\le 1}(x)\}]^\star\,(y) = \|y\|_{\bar{q}} \tag{4.2.2.6}$$

Results are out:

$$h(x) = \|x-b\|_p \quad g(x) = \delta\{\|q\| \le 1\}(x)$$

And using the Rockafella's Duality we have:

$$\sup_{\|y\|_{\bar{p}} \le 1} \left\{ \langle b, y\rangle - \|A^T y\|_{\bar{q}} \right\} \tag{4.2.2.7}$$

The negative sign can get ignore because $g^\star$ is a norm.

4.2.3

The primal is:

$$(P) : \min_x \{\langle c, x\rangle : Ax = b, x \in K\}$$

Where K is a cone.
Identifies the g, h function in this case it's:

$$\begin{aligned} h(Ax) &= \delta\{Ax = b\} \\ \implies h(x) &= \delta\{x = b\} \\ g(x) &= \langle c, x\rangle + \delta\{x \in K\} \end{aligned} \tag{4.2.3.1}$$

We can immdiately identifies the conjugate of the indicator function h because it's part of the Exercise 3.24 from the last homework. And therefore we have: $h^\star(x) = \langle b, x\rangle$.
To figure out the Conjugate of h:

$$\begin{aligned} g^\star(x) &= \sup\{\langle x, y\rangle - g(y)\} \\ &= \sup_y\{\langle x, y\rangle - \langle c, y\rangle - \delta\{y \in K\}(y)\} \\ &= \sup_y\{\langle y, x-c\rangle - \delta\{y \in K\}(y)\} \\ &= \sup_{y \in K}\{\langle y, x-c\rangle\} \\ &= \delta_K^\star(x-c) \end{aligned} \tag{4.2.3.2}$$

Using page 60 of the textbook, and the fact that K is a cone, we have: $g^\star(x) = \delta_K^\star(x-c) = \delta_{K^\circ}(x-c)$.
The Fenchel Dual is:

$$\begin{aligned} \sup_y \{-h^\star(y) - g^\star(-A^* y)\} &= \sup_y \left\{-\langle b, y\rangle - \delta_{K^\circ}(-A^T y - c)\right\} \\ &\equiv \sup_y \left\{\langle b, y\rangle - \delta_{K^\circ}(A^T y - c)\right\} \end{aligned} \tag{4.2.3.3}$$

4.2.4

The Primal Problem is:

$$(P) : \min_x \left\{ \frac{1}{2}\langle x, Qx\rangle + \langle c, x\rangle : Ax \le b \right\}$$

Where the matrix Q is Positive Definite. Identifying the g, h we have:

$$\begin{aligned} h(Ax) &= \delta\{Ax \ge b\}(x) \\ g(x) &= \frac{1}{2}\|x\|_Q^2 + \langle x, c\rangle \\ \implies h(x) &= \delta\{x \ge b\}(x) \\ &= \delta\{x - b \ge 0\}(x) \end{aligned} \tag{4.2.4.1}$$

Notice that $x - b \ge 0 \iff x - b \in \mathbb{R}^n_+$ and $\mathbb{R}^n_+$ is a cone and it's polar cone is $\mathbb{R}^n_-$. Therefore, we use table 3.4:

$$\begin{aligned} h(x) &= \delta_{\mathbb{R}^n_+}(x - b) \\ \implies h(x + b) &= \delta_{\mathbb{R}^n_+}(x) \\ \implies h^\star(x) &= \delta_{\mathbb{R}^n_-}(x) - \langle b, x\rangle \end{aligned} \tag{4.2.4.2}$$

The conjugate of g is given by:

$$\begin{aligned} g^\star(y) &= \left[\frac{1}{2}\|x\|_Q^2 + \langle c, x\rangle \middle| x\right]^\star(y) \\ &= \left[\frac{1}{2}\|x\|_Q^2 \middle| x\right]^\star(y - c) \end{aligned} \tag{4.2.4.3}$$

Let's figure out the conjugate of the one half of the Energy Norm squared:

$$\left[\frac{1}{2}\|A^{1/2}x\|_2^2\right]^\star(y) = \sup_x\left\{\langle y, x\rangle - \frac{1}{2}\|A^{1/2}x\|_2^2\right\} \tag{4.2.4.4}$$

$$\begin{aligned} &\text{Set Gradient to zero to find the sup} \\ \nabla\left[\langle y, x\rangle - \frac{1}{2}\|x\|_A^2 \middle| x\right] &= 0 \\ \implies y - Ax^+ &= 0 \\ Ax^+ &= y \\ x^+ &= A^{-1}y \\ \implies \sup_x\left\{\langle y, x\rangle - \frac{1}{2}\|A^{1/2}x\|_2^2\right\} &= \langle y, A^{-1}y\rangle - \frac{1}{2}\|A^{-1/2}y\|_2^2 \\ &= \|y\|_{A^{-1}}^2 - \frac{1}{2}\|y\|_{A^{-1}}^2 \\ &= \frac{1}{2}\|y\|_{A^{-1}}^2 \end{aligned}$$

Then we can conclude together with the results from (4.2.4.3) to get:

$$g^\star(y) = \frac{1}{2}\|y - c\|_{Q^{-1}}^2 \tag{4.2.4.5}$$

Therefore, the Dual is:

$$\begin{aligned} \max_y\{-h^\star(y) - g^\star(-\mathcal{A}^* y)\} &= \max_y\left\{\delta_{\mathbb{R}^n_-}(y) - \langle b, y\rangle - \frac{1}{2}\| - A^T y - c\|_{Q^{-1}}^2\right\} \\ &\equiv \max_{y\ge 0}\left\{\langle b, y\rangle - \frac{1}{2}\|A^T y - c\|_{Q^{-1}}^2\right\} \end{aligned} \tag{4.2.4.6}$$

5.14

For this part of the HW I implemented the Smooth Gradient Descend and the Accelerated Gradient on Julia the programming language and made the plot of for the values of the objective functions for the first 100 iterations of both method. (Running the code requires installing UnicodePlots.jl and Plots.jl)
This is how I got the gradient for the function for the implementations.
Let $X \in \mathbb{R}^{m\times n}$, let $y \in \mathbb{R}^m$, $\theta \in \mathbb{R}^n$. And let lower case x_i denotes the i th of the matrix X The Gradient

of the loss function is computed as:

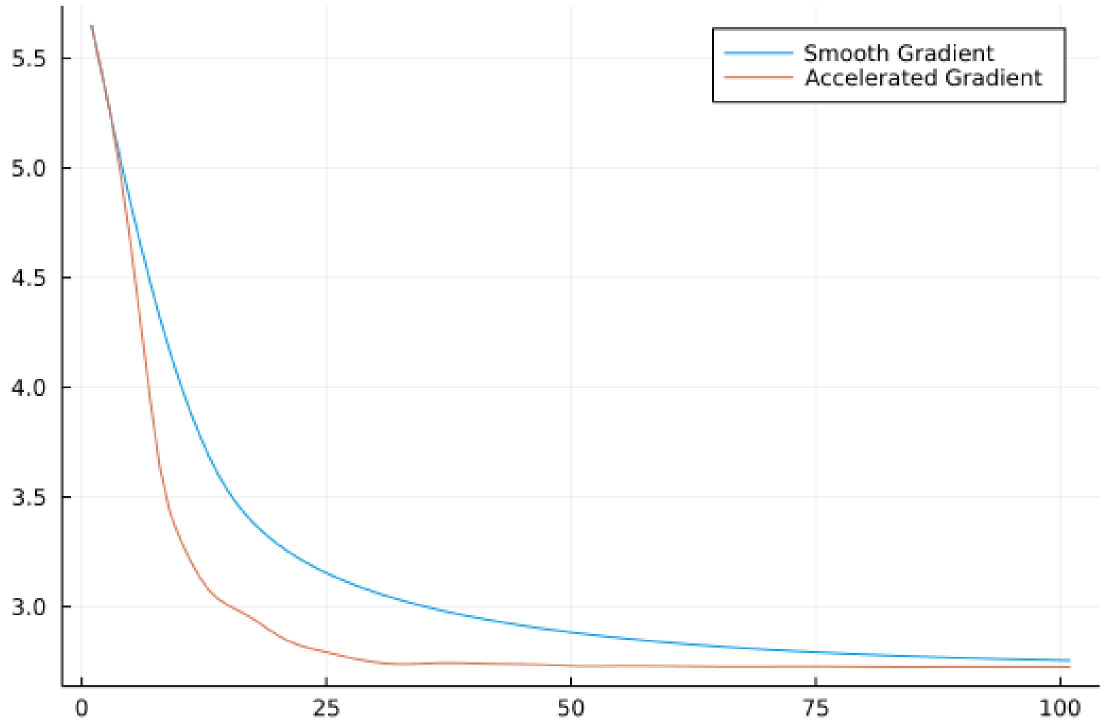
$$\begin{aligned}
&\nabla_\theta\left[\sum_{i=1}^{m}\ln(1+\exp\left(\langle\theta,-y_ix_i\rangle\right))+\frac{\lambda}{2}\|\theta\|_2^2\right] \qquad (5.14.1)\\
&=\left(\sum_{i=1}^{m}\nabla_\theta\left[\ln\left(1+\exp\langle\theta,-y_ix_i\rangle\right)\right]\right)+\lambda\theta\\
&=\left(\sum_{i=1}^{m}\frac{\exp(\theta,-yx_i)}{1+\exp(\theta,-yx_i)}\nabla_\theta[\langle\theta,-y_ix_i\rangle]\right)+\lambda\theta\\
&=\left(\sum_{i=1}^{m}\frac{-\exp(\theta,-yx_i)}{1+\exp(\theta,-yx_i)}yx_i\right)+\lambda\theta
\end{aligned}$$

The code implementations is here:

```
using LinearAlgebra
import UnicodePlots
import Plots
import Logging

# ------------------------------------------------------------------------------
mutable struct LogisticRegressionModel
    X::Matrix{Float64}        # The data
    y::Vector{Float64}        # the labels
    theta::Vector{Float64}    # The parameters
    lambda::Float64           # a regularization parameters

    function LogisticRegressionModel(X, y, theta; lambda=1)
        this = new(X, y, theta, lambda)
        return this
    end

end

function Base.copy(this::LogisticRegressionModel)
    return LogisticRegressionModel(
        copy(this.X),
        copy(this.y),
        copy(this.theta),
        lambda = copy(this.lambda)
        )
end

function ErrorOf(this::LogisticRegressionModel)
    theta = this.theta
    X = this.X
    y = this.y
    lambda = this.lambda

    Error = - y.*(X*theta)  # vector
    Error = 1 .+ exp.(Error)  # vector
    Error .= log.(Error)  # vector
    Error = sum(Error)
    Error += (lambda/2)*dot(theta, theta)
    return Error
end

function GradientOf(this::LogisticRegressionModel)
    theta = this.theta
    X = this.X
    y = this.y
    lambda = this.lambda

    Expr1 = (x) -> exp(x)/(1 + exp(x))
    Summation = zeros(size(theta))
    for Index = 1:size(X, 1)
        x = X[Index, :]
        Expr2 = - y[Index]*x  # vector
        Expr3 = dot(theta, Expr2)  # scalar
        Summation += Expr1(Expr3) * Expr2  # Vector
    end
    # ---------------------------------------------
    # Summation = exp.(-y.*X*theta)
    # Summation = Summation./(Summation .+ 1)
    # Summation = X'*(-y.*Summation)
    return Summation + lambda*theta
end

function ParametersOf(this::LogisticRegressionModel)
    return this.theta
end



# ---------------------------------------------------------------------------------
mutable struct GradientDescend
    beta::Float64
    objective::LogisticRegressionModel
    objective_vals::Vector{Float64}

    function GradientDescend(objective::LogisticRegressionModel, beta::Float64)
        this = new(beta, objective, Vector{Float64}())
        push!(this.objective_vals, ErrorOf(this.objective))  # initial error.
        return this
    end
end


function GradientUpdate(this::GradientDescend)
    DeltaTheta = GradientOf(this.objective)
    Params = ParametersOf(this.objective)
    Params .-= DeltaTheta * (1/this.beta)
    push!(this.objective_vals, ErrorOf(this.objective))

    return norm(DeltaTheta)
end

# ---------------------------------------------------------------------------------
```

```
mutable struct AcceleratedGradientDescend
    beta::Float64
    objective::LogisticRegressionModel
    objective_vals::Vector{Float64}
    t::Int64                    # Iteration number.
    a::Dict{Int64, Float64}
    x::Dict{Int64, Vector{Float64}}

    function AcceleratedGradientDescend(objective::LogisticRegressionModel, beta::Float64)
        this = new(
            beta,
            objective,
            Vector{Float64}(),
            0,
            Dict{Int64, Float64}(),
            Dict{Int64, Vector{Float64}}()
        )
        push!(this.objective_vals, ErrorOf(this.objective))  # initial error.
        a = this.a
        x = this.x
        a[-1] = a[0] = 1
        x[-1] = x[0] = this.objective.theta
        return this
    end
end

function GradientUpdate(this::AcceleratedGradientDescend)
    t = this.t
    x = this.x
    a = this.a
    beta = this.beta
    u = x[t] + a[t]*(a[t - 1]^(-1) - 1)*(x[t] - x[t - 1])
    Gradient = GradientOf(this.objective)
    x[t + 1] = u - (1/beta)*Gradient
    this.objective.theta .= x[t + 1]
    a[t + 1] = (sqrt(a[t]^4 + 4a[t]^2) - a[t]^2)/2
    this.t += 1
    push!(this.objective_vals, ErrorOf(this.objective))
    return norm(Gradient)
end



function Run()
    function Bernoulli(p)
        if rand() < p
            return 1
        end
        return -1
    end
    m, n = 100, 50
    beta = 100.0
    theta = ones(n)
    X = randn(m, n)
    z = X*theta
    p = 1 ./ (1 .+ exp.(-z))
    theta0 = randn(n)
    y = Bernoulli.(p)
    Objective = LogisticRegressionModel(X, y, theta0)
    Optim = GradientDescend(Objective, beta)
    OptimAcc = AcceleratedGradientDescend(copy(Objective), beta)
    println("running...")
    for _ in 1:100
        println(GradientUpdate(Optim))
        GradientUpdate(OptimAcc)
    end
    Logging.@info "Objective Values (Smooth Gradient Descend):"
    FxnVals = Optim.objective_vals
    FxnValsAcc = OptimAcc.objective_vals
    display(FxnVals)
    Logging.@info "Objective Values (Acc Gradient Descend)"
    display(FxnValsAcc)

    # Plot this out -------------------------------------------------------------
    UPlot = UnicodePlots.lineplot(1:length(FxnVals), FxnVals)
    println(UnicodePlots.lineplot!(UPlot, 1:length(FxnValsAcc), FxnValsAcc))
    Plots.plot(collect(1:length(FxnVals)), log.(FxnVals), label="Smooth Gradient")
    Plots.plot!(collect(1:length(FxnValsAcc)), log.(FxnValsAcc), label="Accelerated Gradien

    Plots.savefig("objectiveVals.png")

    return Optim
end

Optim = Run()


```

The plot produced by the algorithms is here(Next page):



Hi Chatherine, if you need the file for the code and the Julia Project Environment, please contact me I am happy to provide.

5.15

5.15.1

We assume that the $f(x)$ is convex differentiable with minimizer $x^\star$. And we wish to show the following:

$$\text{Gradient Update: } x_{k+1} = x_k + \lambda_k \nabla f(x_k) \tag{5.15.1.1}$$

$$\text{Wish to prove: } \frac{1}{2}\|x_{k+1} - x_*\|^2 \le \frac{1}{2}\|x - x_*\|^2 - \gamma_k(f(x_k) - f(x_*)) + \frac{\gamma_k}{2}\|\nabla f(x_k)\|$$

To start consider that:

$$\begin{aligned}\|(x_{k+1} - x_k) - (x_k - x_*)\|^2 &= \|x_{k+1}\|^2 + \|x_k - x_*\|^2 + 2\langle x_{k+1} - x_k, x_k - x_*\rangle \\ &= \|x - x_*\|^2 + \|\gamma_k \nabla f(x_k)\|^2 + 2\langle \gamma_k \nabla f(x_k), x_k - x_*\rangle\end{aligned} \tag{5.15.1.2}$$

Divides the whole expression by 2 and focuses on the last 2 terms of the expression:

$$\begin{aligned}&\frac{1}{2}\|\gamma_k \nabla f(x_k)\|^2 + \langle \gamma_k \nabla f(x_k), x_k - x_*\rangle \\ =&\frac{\gamma_k^2}{2}\|\nabla f(x_k)\|^2 + \gamma_k\langle \nabla f(x_k), x_k - x_*\rangle\end{aligned} \tag{5.15.1.3}$$

Using Excercise 4.12 #4 we have:

$$\begin{aligned}\langle \nabla f(x_k) - \underbrace{\nabla f(x_*)}_{=0}, x_* - x_k\rangle \ge 0 \\ f(x_*) - f(x_k) \ge \langle \nabla f(x_k), x_k - x_*\rangle\end{aligned} \tag{5.15.1.4}$$

Apply (5.15.1.4) to (5.15.1.3) we have:

$$\frac{\gamma_k^2}{2}\|\nabla f(x_k)\|^2 + \gamma_k\langle \nabla f(x_k), x_k - x_*\rangle \le \frac{\gamma_k^2}{2}\|\nabla f(x_k)\|^2 + \gamma_k(f(x_*) - f(x_k)) \tag{5.15.1.5}$$

Go back to (5.15.1.2) and apply that previous bound we have:

$$\begin{aligned}\frac{1}{2}\|(x_{k+1} - x_k) - (x_k - x_*)\|^2 &= \frac{1}{2}\|x - x_*\|^2 + \frac{1}{2}\|\gamma_k\nabla f(x_k)\|^2 + \langle \gamma_k\nabla f(x_k), x_k - x_*\rangle \\ &\le \frac{1}{2}\|x - x_*\|^2 + \frac{\gamma_k^2}{2}\|\nabla f(x_k)\|^2 + \gamma_k(f(x_*) - f(x_k)) \\ &= \frac{1}{2}\|x - x_*\|^2 + \frac{\gamma_k^2}{2}\|\nabla f(x_k)\|^2 - \gamma_k(f(x_k) - f(x_*))\end{aligned} \tag{5.15.1.6}$$

Which is exactly what we want to show.

5.15.2

We wish to show that that sequence is minimized by choosing $\gamma_k = (f(x_k) - f^*)/\|\nabla f(x_k)\|^2$, where f^* is the minimal value for the function. This is what we wish to prove:

$$\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 - \left(\frac{f(x_k) - f^*}{\|\nabla f(x_k)\|}\right)^2$$

And then proving the the sequence is bounded by:

$$\begin{aligned}\frac{1}{2}\|x_{k+1} - x_*\|^2 &\le \frac{1}{2}\|x_k - x_*\|^2 - \gamma_k(f(x_k) - f^*) + \frac{\gamma_k^2}{2}\|\nabla f(x_k)\|^2 \\ &= \frac{1}{2}\|x_k - x_*\|^2 - \gamma_k\left(f(x_k) - f^* - \frac{\gamma_k}{2}\|\nabla f(x_k)\|^2\right) \\ \text{Let: } &\partial_{\gamma_k}\left[\gamma_k(f(x_k) - f^*) + \frac{\gamma^2}{2}\|\nabla f(x_k)\|^2\right] = 0 \\ -(f(x_k) - f^*) + \gamma_k^+\|\nabla f(x_k)\|^2 = 0 & \\ \gamma_k^+ &= \frac{f(x_k) - f^*}{\|\nabla f(x_k)\|^2}\end{aligned} \tag{5.15.2.1}$$

Consider the RHS of the first expression in the above block of statements, and substitute the optimal stepside of the expression we have:

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &\le \|x_k - x_*\|^2 - 2\gamma_k^+(f(x_k) - f^*) + (\gamma_k^+)^2\|\nabla f(x_k)\|^2 \\ &= \|x_k - x_*\|^2 - \gamma_k^+(2(f(x_k) - f^*) - \gamma^+\|\nabla f(x_k)\|^2) \\ &= \|x_k - x_*\|^2 - \gamma_k^+(2(f(x_k) - f^*) - (f(x_k) - f^*)) \\ &= \|x_k - x_*\| - \gamma_k^+(f(x_k) - f^*) \\ &= \|x_k - x_*\| + \left(\frac{f(x_k) - f^*}{\|\nabla f(x_k)\|}\right)^2\end{aligned} \tag{5.15.2.2}$$

Which is the statement that we wish to prove.

5.15.3

In this part we wish to prove that :

$$f\left(\frac{1}{k}\sum_{i=0}^{k-1} x_i\right) - f^* \le \frac{2\beta\|x_0 - x_*\|^2}{k}$$

Inaddition when the function is α strongly convex, wehave the additional results of:

$$\|x_{k+1} - x_*\|^2 \le \left(1 - \frac{\alpha}{4\beta}\right)\|x_k - x_*\|^2$$

Proof: To make our life easier, we consider using the following notations for the errors and the optimality gap:

$$\begin{aligned} E_k &= f(x_k) - f^* \\ e_k &= x_k - x^* \end{aligned} \tag{5.15.3.1}$$

Starting with exercise 3.12, the differential characterization for beta smooth function, we have:

$$\begin{aligned} f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 &\le f(y) \\ &\text{consider: } y = x_k \quad x = x^* \\ f^* + \frac{1}{2\beta}\|\nabla f(x_k)\|^2 &\le f(x_k) \\ &\|\nabla f(x_k)\|^2 \le 2\beta(f(x_k) - f^*) = 2\beta E_k \\ &\implies \frac{1}{2\beta} \le \frac{E_k}{\|\nabla f(x_k)\|^2} \\ &\implies \frac{E_k}{2\beta} = \frac{E_k^2}{\|\nabla f(x_k)\|^2} \end{aligned} \tag{5.15.3.2}$$

Then we start with the results from what we proved in (5.15.2.2), And written it as:

$$\begin{aligned} \|e_{k+1}\|^2 &\le \|e_k\|^2 - \left(\frac{E_k}{\|\nabla f(x_k)\|}\right)^2 \\ \left(\frac{E_k}{\|\nabla f(x_k)\|}\right)^2 &\le \|e_k\|^2 - \|e_{k+1}\|^2 \\ \underset{(5.15.3.2)}{\implies} \frac{E_k}{2\beta} &\le \|e_k\|^2 - \|e_{k+1}\|^2 \end{aligned} \tag{5.15.3.3}$$

Consider summing up the terms in the above inequalities for $0 \le i \le k-1$:

$$\begin{aligned} \sum_{i=0}^{k-1}\frac{E_i}{2\beta} &\le \sum_{i=1}^{k-1}\|e_i\|^2 - \|e_{i+1}\|^2 \\ \frac{1}{k}\sum_{i=1}^{k-1}E_i &\le \frac{2\beta}{k}\sum_{i=1}^{k-1}\|e_i\|^2 - \|e_{i+1}\|^2 \\ \frac{1}{k}\sum_{i=0}^{k-1}E_i &\le \frac{2\beta}{k}\left(\|e_0\|^2 - \|e_k\|^2\right) \\ \underset{\text{By Convexity}}{\implies} f\left(\frac{1}{k}\sum_{i=0}^{k-1}E_i\right) - f^* &\le \frac{2\beta}{k}\left(\|e_0\|^2 - \|e_k\|^2\right) \\ f\left(\frac{1}{k}\sum_{i=0}^{k-1}E_i\right) - f^* &\le \frac{2\beta}{k}\|e_0\|^2 \\ f\left(\frac{1}{k}\sum_{i=0}^{k-1}x_i\right) - f^* &\le \frac{2\beta\|x_0 - x_*\|^2}{k} \end{aligned} \tag{5.15.3.4}$$

I used telescoping from the second line to the first line in the above block. Just expand the sume and the adjacent terms of the differences will get canceled out.

From the second last line to the last line, we can remove the $-\|e_k\|^2$ becase it's larger than zero and removing it will make the quantity larger.

Now let's consider the statement from 3.57, the differential characterization of α strongly convex function:

$$f(x_k) - f^* \geq \frac{\alpha}{2}\|x_k - x^*\|^2 \tag{5.15.3.5}$$

$$\|\nabla f(x_k)\| = \|\nabla f(x_k) - \underbrace{\nabla f(x_*)}_{=0}\| \leq \beta\|x_k - x^*\| \tag{5.15.3.6}$$

Now we may reconsider the results from (5.15.3.4), which is giving us:

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - \left(\frac{f(x-k) - f^*}{\|\nabla f(x_k)\|}\right) \qquad (5.15.3.7) \\ \|x_{k+1} - x_*\|^2 &\leq \|x_k - x^*\|^2 - \left(\frac{\frac{\alpha}{2}\|x_k - x^*\|^2}{\beta\|x_k - x_*\|^2}\right) \\ &= \left(1 - \frac{\alpha^2}{4\beta^2}\right)\|x_k - x_*\|^2 \end{aligned}$$

From the first expression to the second one, I replaces the numerator with a smaller quantity, the denominator with a larger quantity, reducing the size of the fraction as whole, and hence reducing the quantity being subtracted smaller, placing an upper bound on the RHS of the first inequality.

That last expression is what we wish to prove for the problem.

6.11

We were given the accelerated Gradient Descend method which has an interesting sequence that weights on the previous velocity of the descend steps.

6.11.1

We wish to show that this relation for the sequence holds:

$$\frac{1 - a_{t+1}}{a_{t+1}^2} = \frac{1}{a_t^2} \quad \forall t \geq 0 \tag{6.11.1.1}$$

Proof:

$$\begin{aligned} a_{t+1} &= \frac{1}{2}\left(\sqrt{a_t^4 + 4a_t^2} - a_t^2\right) \qquad (6.11.1.2) \\ 2a_{t+1} &= \sqrt{a_t^4 + 4a_t^2} - a_t^2 \\ 2a_{t+1} + a_t &= \sqrt{a_t^4 + 4a_t^2} \\ (2a_{t+1} + a_t^2)^2 &= a_t^4 + 4a_t^2 \\ 4a_{t+1}^2 + a_t^4 + 4a_t^2 a_{t+1} &= a_t^4 + 4a_t^2 \\ 4a_{t+1}^2 + 4a_{t+1}a_t^2 - 4a_t^2 &= 0 \\ a_{t+1}^2 + a_{t+1}a_t^2 - a_t^2 &= 0 \\ a_{t+1}^2 + a_t^2(a_{t+1} - 1) &= 0 \\ \frac{1}{a_t^2} + \frac{a_{t+1} - 1}{a_{t+1}^2} &= \frac{0}{a_t^2 a_{t+1}^2} \\ \frac{a_{t+1} - 1}{a_{t+1}^2} &= \frac{1}{a_t^2} \quad \forall t \geq 0 \end{aligned}$$

6.11.2

We wish to prove that $\sum_{i=0}^{t}\frac{1}{a_i} = \frac{1}{a_t^2}$, and then the sequence a_k is upper bounded by the sequence $2/(t+2)$. Inductively assume that the statement holds for n:

$$\sum_{i=0}^{t}\frac{1}{a_i} = \frac{1}{a_t^2} \quad \forall t \le n \tag{6.11.2.1}$$

$$\begin{aligned}
\frac{1-a_{t+1}}{a_{t+1}^2} &= \frac{1}{a_t^2} \\
\frac{1}{a_{t+1}^2} - \frac{1}{a_{t+1}} &= \frac{1}{a_t^2} \\
\frac{1}{a_{t+1}^2} &= \frac{1}{a_t^2} + \frac{1}{a_{t+1}} \\
\frac{1}{a_{t+1}^2} &= \sum_{i=0}^{t} a_i + \frac{1}{a_{t+1}} \\
\frac{1}{a_{t+1}^2} &= \sum_{i=0}^{t+1}\frac{1}{a_i}
\end{aligned} \tag{6.11.2.2}$$

The statement paramaterized by $n+1$ holds true. The base case it's not hard to verify, just compute:

$$\begin{aligned}
\frac{1}{a_1} + \frac{1}{a_0} &= \frac{1}{\frac{1}{2}(\sqrt{5}-1)} + 1 \\
\frac{1}{a_1^2} &= \frac{1}{\frac{1}{4}(\sqrt{5}-1)^2}
\end{aligned} \tag{6.11.2.3}$$

Which is actually proved by the end of (6.11.1.2).
Next, we wish to prove the upper bound for the sequence.

This problem is too hard and I can't prove it. I accessed the solutions to multiple peers and they turned out to contain their own algebraic errors. I don't think the upper bound is proven via induction. I do know that the sequence approaches zero monotone. And judging by the amount of efforts I put into this problem, I don't think it's worth it.

Inductively, I used a lot of methods and the fact that the sequence is monotone decreasing.

I gave up after discovering something huge. I try to insert epsilons in inequalities trying to look for additional assumptions can make it better, however, it's futile because all epsilon has to be less than zero, proving that all inductive proof is likely to fail.

Therefore, the proof for the question is not via induction, and it's very likely that this is true.

I computationally verify that the upper bound is extremely tight, and the difference between the sequence $2/(k+2)$ and a_k is a power sequence of power -2. The bound is very tight.

Hi Catherine, please check other's algebra on this problem carefully, the proof you went through on Thursday from a student has an algebraic mistake. If you find any legit proof for this problem, please share it via Canvas message to me, I want to know how to handle such a tight bound for this sequence.

Here is the code I used for the numerical investigation of the problem. You will need it to intuitively verify the correct solutions. The code is in matlab.

```
close all;
a = 1;
Points = 0:300;
for Index = Points(2:end)
    p = a(end); % Previous
    a(end + 1) = (sqrt(p^4 + 4*p^2) - p^2)/2;
end
figure;
plot(a, ’-o’, "linewidth", 1); hold on
UpperBound = 2./(Points + 2);
LowerBound = 2./(Points + 3);
plot(UpperBound, "linewidth", 1);
plot(LowerBound, "linewidth", 1);
legend(["sequence", "upper bound", "Lower Bound"])

figure;
DiffBand = UpperBound - LowerBound;
loglog(DiffBand);
Coeff = polyfit(log(Points(2:end)), log(DiffBand(2:end)), 1);
disp("Log log Coefficient:")
disp(num2str(Coeff(1)))
```