Name: □

Roll No.: □ **Dept.:** □

IIT Kanpur
CS777 Stat. Alg. Learning Theory
End-semester Examination
Date: April 22, 2018

Instructions: *Total:* **100 marks**

1. This question paper contains a total of 8 pages (8 sides of paper). Please verify.
2. Write your name, roll number, department on **every side of every sheet** of this booklet.
3. Write final answers **neatly with a pen**. Pencil marks can get smudged and you may lose credit.
4. Do not give derivations/elaborate steps unless the question specifically asks you to provide these.
5. You may use results proved in class directly. However, clearly state the result being used.

Problem 1 (True or False: 10 X 1 = 10 marks). For each of the following simply write **T** or **F** in the box.

1. $\boxed{\text{F}}$ The majorization minimization algorithm provides monotonic progress i.e. $f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t)$ only when the objective function f is convex. (Monotonic progress is guaranteed universally.)

2. $\boxed{\text{T}}$ A function that satisfies α-strong convexity over $\mathbb{R}^d$ for some $\alpha > 0$ always satisfies $\alpha/2$-restricted strong convexity over s-sparse vectors in $\mathbb{R}^d$. (A strongly convex function retains this property upon restriction. RSC is a monotonic property α-RSC $\Rightarrow$ α'-RSC for all $\alpha' \leq \alpha$.).

3. $\boxed{\text{T}}$ The most disk-access friendly variant of coordinate descent is cyclic coordinate descent that chooses coordinates in lexicographic order i.e. $1, 2, \ldots, d, 1, 2, \ldots, d, 1, 2, \ldots$. (Yes, lexicographic orders can ensure fast disk access and higher success rates for speculative caching algorithms.)

4. $\boxed{\text{F}}$ The Fenchel dual of a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex only when the function f is itself convex over the entire space $\mathbb{R}^d$. (The Fenchel dual function is universally convex.)

5. $\boxed{\text{T}}$ If we have two families of binary classifiers $\mathcal{F}, \mathcal{F}' \subset \{-1, +1\}^{\mathcal{X}}$ over a domain $\mathcal{X}$ such that $\mathcal{F} \subseteq \mathcal{F}'$, then VC-dim$(\mathcal{F}) \leq$ VC-dim$(\mathcal{F}')$. (Yes, whenever a point set is shattered by $\mathcal{F}$, it is definitely shattered by $\mathcal{F}'$.)

6. $\boxed{\text{T}}$ Suppose we have a transition law $P \in \mathbb{R}^{N \times N}$ over a finite domain $[N]$ such that $P_{i,j} > 0$ for all $i, j \in [N]$. Such a law P is always aperiodic and irreducible. (Yes, P corresponds to a "complete" graph which mixes in constant time independent of its size.)

7. $\boxed{\text{F}}$ The Kullback-Leibler divergence between any two distributions is always bounded between 0 and 1. (No, consider the KL-divergence between two Bernoulli distributions $KL(p\|q)$ where $p \approx 0.5$ and $q \to 0$.)

8. $\boxed{\text{F}}$ The set of right stochastic matrices $P \in \mathbb{R}^{N \times N}$ over a finite domain $[N]$ is closed under matrix addition. (False, however, they do form a convex set.)

9. $\boxed{\text{T}}$ The set of right stochastic matrices $P \in \mathbb{R}^{N \times N}$ over a finite domain $[N]$ is closed under matrix multiplication. (Yes, $PQ\mathbf{1} = P\mathbf{1} = \mathbf{1}$.)

10. $\boxed{\text{F}}$ Alternating minimization always converges to the global optimum for all convex functions. (No, for non-smooth functions, it can get stuck. For example, consider $f(x, y) = \max\{|x|, |y|\}$. The function is clearly convex and the optimum is at $(0, 0)$ but if we initialize it at $(1, 1)$, then AM remains stuck there since $f(1, |y|) = 1 = f(1, 1)$ for all $|y| \leq 1$ and $f(1, |y|) > 1$ for all $|y| > 1$.)

Problem 2 (Short Questions: 5 x 5 = 25 marks). Answer questions in brief in the space given below.

1. Strong smoothness does not imply convexity. Find a non-convex function that is 1-strongly smooth.

   Solution: Consider $f(x) = -\frac{1}{2}x^2$ i.e. the inverted parabola. It is concave and decidedly non-convex. However, it satisfies $f(x) = -\frac{1}{2}x^2 \leq -\frac{1}{2}x^2 + (x-y)^2 = -\frac{1}{2}y^2 - y(x-y) + \frac{1}{2}(x-y)^2 = f(y) + f'(y) \cdot (x-y) + \frac{1}{2}(x-y)^2$. Thus, f is smooth but non-convex.

2. Show that the set of s-sparse vectors in d dimensions is non-convex for any $s < d$.

   Solution: For any $1 \leq s < d$ let $\mathbf{a} = \sum_{i=1}^{s} \mathbf{e}_i$ and $\mathbf{b} = \sum_{j=2}^{s+1} \mathbf{e}_j$ where $\mathbf{e}_i$ is the i^{th} canonical vector with 1 at the i^{th} location and zero everywhere else. Clearly $\mathbf{a}, \mathbf{b}$ are both s-sparse. However $(\mathbf{a}+\mathbf{b})/2 = \sum_{i=2}^{s} \mathbf{e}_i + 0.5 \cdot \mathbf{e}_1 + 0.5 \cdot \mathbf{e}_{s+1}$ which has $(s+1)$ non-zero coordinates. Thus, the set of s-sparse vectors is non-convex.

3. Give Coordinate Minimization updates for $\min_{\mathbf{x} \in \mathbb{R}^d} \frac{1}{2}\left\|\mathbf{b} - A\mathbf{x}\right\|_2^2 + \frac{1}{2}\left\|\mathbf{x}\right\|_2^2$ where $A \in \mathbb{R}^{n \times d}, \mathbf{b} \in \mathbb{R}^n$.

   Solution: Let $A_i \in \mathbb{R}^n$ denote the i^{th} column of the matrix A. Thus $\mathbf{b} - A\mathbf{x} = \mathbf{b} - \sum_{i=1}^{d} \mathbf{x}_i \cdot A_i := L_j(\mathbf{x}) - \mathbf{x}_j \cdot A_j$ where $L_j(\mathbf{x}) = \mathbf{b} - \sum_{i \neq j}^{d} \mathbf{x}_i \cdot A_i$. Also let $R_j(\mathbf{x}) := \sum_{i \neq j}^{d} \mathbf{x}_i^2$. Note that L_j, R_j do not depend on the value of $\mathbf{x}_j$ at all. Thus, we have

   $$\frac{1}{2}\left\|\mathbf{b} - A\mathbf{x}\right\|_2^2 + \frac{1}{2}\left\|\mathbf{x}\right\|_2^2 = \frac{1}{2}\left\|L_j(\mathbf{x}) - \mathbf{x}_j \cdot A_j\right\|_2^2 + \frac{1}{2}(\mathbf{x}_i^2 + R_j(\mathbf{x}))$$

   The portions of the above expression that actually depend on $\mathbf{x}_j$ are

   $$\frac{1}{2}\left\|A_j\right\|_2^2 \cdot \mathbf{x}_j^2 - \cdot \mathbf{x}_j \cdot \langle A_j, L_j(\mathbf{x}) \rangle + \frac{1}{2}\mathbf{x}_i^2$$

   Minimizing the above over $\mathbf{x}_j$ gives us

   $$\mathbf{x}_j = \frac{\langle A_j, L_j(\mathbf{x}) \rangle}{1 + \left\|A_j\right\|_2^2}$$

4. For any (possibly non-convex) function $f : \mathbb{R}^d \to \mathbb{R}$, show that $f^{**}(\mathbf{x}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$. Hint: first show that for any $\mathbf{z}, \mathbf{x}$, we must always have $f(\mathbf{x}) + f^*(\mathbf{z}) \geq \langle \mathbf{x}, \mathbf{z} \rangle$.

   Solution: We have $f^*(\mathbf{z}) = \max_{\mathbf{x} \in \mathbb{R}^d} \{\langle \mathbf{x}, \mathbf{z} \rangle - f(\mathbf{x})\}$. This means for any $\mathbf{z}, \mathbf{x}$, we must always have $f^*(\mathbf{z}) \geq \langle \mathbf{x}, \mathbf{z} \rangle - f(\mathbf{x})$ i.e. $f(\mathbf{x}) \geq \langle \mathbf{x}, \mathbf{z} \rangle - f^*(\mathbf{z})$. Since this inequality holds for all $\mathbf{z}$, it must hold even if we maximize over $\mathbf{z}$ i.e. $f(\mathbf{x}) \geq \max_{\mathbf{z} \in \mathbb{R}^d} \{\langle \mathbf{x}, \mathbf{z} \rangle - f^*(\mathbf{z})\} =: f^{**}(\mathbf{x})$.

5. Give an example of a coordinate-wise convex function that is not (jointly) convex in all its coordinates.

   Solution: Consider $f(x, y) = xy$. For every fixed value of x, the function is linear in y, hence convex. Similarly, for every fixed value of y, the function is linear in x, hence convex. However $f(1, -1) = -1 = f(-1, 1)$ but $f(0, 0) = 0 \not\leq -1 = (f(1, -1) + f(-1, 1))/2$ so f is not jointly convex.

Problem 3 (Too many coins: 2+5+3 = 10 marks)**.** Consider the problem setting and answer the questions.

```
Algorithm 1: Multi Toss (MCT)

Input: $K$ biases $b_k \in [0.25, 0.75]$
Sample $r \sim \mathsf{UNIF}([0,1])$
for $k = 1, 2, \ldots, K$ do
    if $r > b_k$ then
        Let $X_k \leftarrow 0$
    else
        Let $X_k \leftarrow 1$
    end if
end for
return $X_1, X_2, \ldots, X_K$
```

My friend Jane is running code that requires access to coin tosses with a variety of biases. At each time step, she requires K coin tosses with biases $0.25 \leq b_1 < b_2 < \ldots < b_K \leq 0.75$. She asks me to supply these coin tosses. Now, since K is large, simulating so many coins independently may be expensive so I try to cheat. At every time step, I draw a single random number (independent of previous draws) uniformly from the interval $[0, 1]$ and use Algorithm 1 to simulate the tosses

1. For any fixed $k \in [K]$, what is $\mathbb{P}[X_k = 1]$? Justify your answer.
2. Running MCT T times gives $X_1^t, \ldots X_K^t$ for $t \in [T]$. Let $\hat{X}_k^T = \frac{1}{T}\sum_{t=1}^T X_k^t$. Analyze $\mathbb{P}\left[\hat{X}_k^T > b_k(1+\epsilon)\right]$ for any $k \in [K], \epsilon \in (0,1)$.
3. Analyze $\mathbb{P}\left[\exists k \in [K] \text{ s.t. } \hat{X}_k^T > b_k(1+\epsilon)\right]$

Your answers should be in terms of T, K, ϵ, b_k and universal constants.

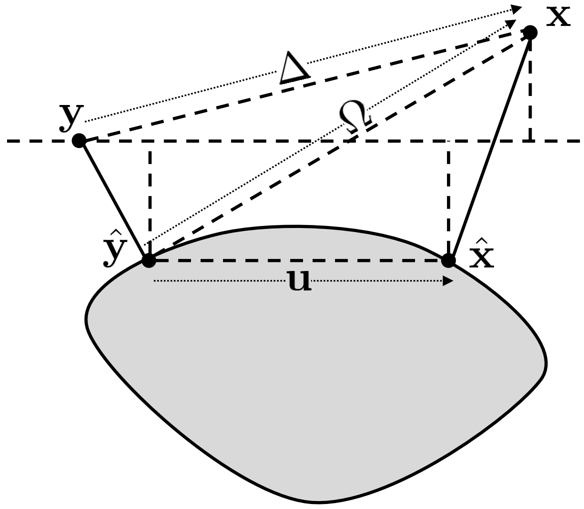
Solution: We give the solution in parts

1. For any k, we have $\mathbb{P}[X_k = 1] = \mathbb{P}[r \leq b_k] = b_k$ since r is drawn from the uniform distribution over $[0, 1]$.

2. Since the random numbers $r^1, \ldots, r^t$ drawn at the different time instants are independent, any function of these would correspond to independent random variables too. In particular $X_k^t = \mathbb{I}\{r^t \leq b_k\}$ are independent variables and, by part 1, identical too. Since these are Boolean, we can apply the Chernoff's inequality to get
$$\mathbb{P}\left[\hat{X}_k^T > b_k(1+\epsilon)\right] \leq \exp(-Tb_k\epsilon^2/3)$$

3. Applying a union bound gives us (since all $b_k \geq 0.25$)
$$\mathbb{P}\left[\exists k \in [K] \text{ s.t. } \hat{X}_k^T > b_k(1+\epsilon)\right] = \mathbb{P}\left[\bigcup_{k=1}^K \left\{\hat{X}_k^T > b_k(1+\epsilon)\right\}\right] \leq \sum_{k=1}^K \mathbb{P}\left[\hat{X}_k^T > b_k(1+\epsilon)\right] \leq K\exp\left(-\frac{T\epsilon^2}{12}\right)$$

Problem 4 (Bringing the World Closer: 25 marks)**.** Consider a convex set $\mathcal{C} \in \mathbb{R}^d$ and any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Let $\hat{\mathbf{x}} = \Pi_\mathcal{C}(\mathbf{x})$ and $\hat{\mathbf{y}} = \Pi_\mathcal{C}(\mathbf{y})$ denote the projections of the two points onto the convex set w.r.t. the Euclidean norm i.e. $\Pi_\mathcal{C}(\mathbf{x}) = \arg\min_{\mathbf{z}\in\mathcal{C}} \left\|\mathbf{z} - \mathbf{x}\right\|_2$. Then prove that $\left\|\hat{\mathbf{x}} - \hat{\mathbf{y}}\right\|_2 \leq \left\|\mathbf{x} - \mathbf{y}\right\|_2$. This shows that convex Euclidean projections are always *contractive*. You may follow the given proof recipe or use your own technique.



If you do decide to follow the proof recipe, you must show the statements in bold explicitly. You may want to attempt parts in the order 1 → 2 → 3 → 5 → 4. If you use the recipe, the marks distribution will be $2 + 4 + 3 + 8 + 8$ otherwise it will be based on your method. Irrespective of your approach, you are free to use the projection properties (O, I and II), after stating them.

Denote $\Delta := \mathbf{x} - \mathbf{y}$, $\mathbf{u} := \hat{\mathbf{x}} - \hat{\mathbf{y}}$ and $\Omega := \mathbf{x} - \hat{\mathbf{y}}$. Hint: to visualize better, imagine $\hat{\mathbf{y}}$ to be the origin.

1. **Show that we can assume that $\left\|\mathbf{u}\right\|_2 = 1$ without loss of generality**.
2. **Give expressions for $\Delta_\parallel$ and $\Omega_\parallel$, the components of Δ and Ω resp. parallel to u. Also give expressions for $\Delta_\perp$ and $\Omega_\perp$, the components of Δ and Ω resp. perpendicular to u**.
3. **Show that $\left\|\boldsymbol{\Delta}\right\|_2^2 \geq \left\|\boldsymbol{\Delta}_\parallel\right\|_2^2$**.
4. **Show that $\left\|\boldsymbol{\Delta}_\parallel\right\|_2^2 \geq \left\|\boldsymbol{\Omega}_\parallel\right\|_2^2$**.
5. **Show that $\left\|\boldsymbol{\Omega}_\parallel\right\|_2^2 \geq \left\|\hat{\mathbf{x}} - \hat{\mathbf{y}}\right\|_2^2$**. This proves that $\left\|\mathbf{x} - \mathbf{y}\right\|_2^2 \geq \left\|\Delta_\parallel\right\|_2^2 \geq \left\|\Omega_\parallel\right\|_2^2 \geq \left\|\hat{\mathbf{x}} - \hat{\mathbf{y}}\right\|_2^2$.

Solution: We give the solution in parts below

1. If $\hat{\mathbf{x}} = \hat{\mathbf{y}}$ i.e. $\mathbf{u} = \mathbf{0}$, then we are already done since we always have $\left\|\mathbf{x} - \mathbf{y}\right\|_2 \geq 0$. Otherwise we shrink/expand the entire space using $\mathbf{x} \mapsto \mu \cdot \mathbf{x}$ with $\mu = \frac{1}{\left\|\mathbf{u}\right\|_2}$ so that we have $\left\|\mathbf{u}\right\|_2 = 1$. Note that this map is linear and preserves order of distances i.e. $\left\|\mu \cdot \mathbf{v}\right\|_2 \geq \left\|\mu \cdot \mathbf{w}\right\|_2$ iff $\left\|\mathbf{v}\right\|_2 \geq \left\|\mathbf{w}\right\|_2$. This map also keeps acute angles acute and obtuse angles obtuse.

2. Let $\Delta_\parallel := \mathbf{u}\mathbf{u}^\top\Delta$ and $\Delta_\perp := (I - \mathbf{u}\mathbf{u}^\top)\Delta$. Notice that $\Delta_\parallel + \Delta_\perp = \Delta$ and $\Delta_\parallel^\top\Delta_\perp = (\mathbf{u}\mathbf{u}^\top - \mathbf{u}\mathbf{u}^\top)\Delta = 0$. Similarly, define $\Omega_\parallel = \mathbf{u}\mathbf{u}^\top\Omega$ and $\Omega_\perp = (I - \mathbf{u}\mathbf{u}^\top)\Omega$.

3. Applying Pythagoras' theorem now gives us $\left\|\Delta\right\|_2^2 \geq \left\|\Delta_\parallel\right\|_2^2$.

4. We have

$$\begin{aligned}\left\|\Delta_\parallel\right\|_2^2 &= (\mathbf{x} - \mathbf{y})^\top\mathbf{u}\mathbf{u}^\top(\mathbf{x} - \mathbf{y}) = (\mathbf{x} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \mathbf{y})^\top\mathbf{u}\mathbf{u}^\top(\mathbf{x} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \mathbf{y})\\ &= \left\|\Omega_\parallel\right\|_2^2 + \left\|\mathbf{u}\mathbf{u}^\top(\hat{\mathbf{y}} - \mathbf{y})\right\|_2^2 + 2(\hat{\mathbf{y}} - \mathbf{y})^\top\mathbf{u}\mathbf{u}^\top(\mathbf{x} - \hat{\mathbf{y}})\end{aligned}$$

   Applying projection property-I tells us that $(\hat{\mathbf{y}} - \mathbf{y})^\top\mathbf{u} \geq 0$ whereas $\mathbf{u}^\top(\mathbf{x} - \hat{\mathbf{y}}) = (\hat{\mathbf{x}} - \hat{\mathbf{y}})^\top(\mathbf{x} - \hat{\mathbf{y}}) = \left\|\hat{\mathbf{x}} - \hat{\mathbf{y}}\right\|_2^2 + (\hat{\mathbf{x}} - \hat{\mathbf{y}})^\top(\mathbf{x} - \hat{\mathbf{x}})$. However $(\hat{\mathbf{x}} - \hat{\mathbf{y}})^\top(\mathbf{x} - \hat{\mathbf{x}}) \geq 0$ by projection property-I which gives us $\mathbf{u}^\top(\mathbf{x} - \hat{\mathbf{y}}) \geq 0$. This proves the claim $\left\|\Delta_\parallel\right\|_2^2 \geq \left\|\Omega_\parallel\right\|_2^2$.

5. We have $\left\|\Omega_\parallel\right\|_2^2 = (\mathbf{x} - \hat{\mathbf{y}})^\top\mathbf{u}\mathbf{u}^\top(\mathbf{x} - \hat{\mathbf{y}}) = (\mathbf{x} - \hat{\mathbf{x}} + \mathbf{u})^\top\mathbf{u}\mathbf{u}^\top(\mathbf{x} - \hat{\mathbf{x}} + \mathbf{u}) = 1 + ((\mathbf{x} - \hat{\mathbf{x}})^\top\mathbf{u})^2 + 2(\mathbf{x} - \hat{\mathbf{x}})^\top\mathbf{u}$. Applying projection property-I tells us that $(\mathbf{x} - \hat{\mathbf{x}})^\top\mathbf{u} \geq 0$ which gives us $\left\|\Omega_\parallel\right\|_2^2 \geq 1 = \left\|\hat{\mathbf{x}} - \hat{\mathbf{y}}\right\|_2^2$.

```
Algorithm 2: Pseudoconvex Programming (PCP)

Input: Convex function $f$, Concave function $g$
Let $\mathbf{x}^0 \in \mathcal{C}$                          //Initialize
for $t = 1, 2, \ldots, T-1$ do
    Let $\lambda_t \leftarrow h(\mathbf{x}^t)$
    Let $\mathbf{x}^{t+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{C}} \ V_{\lambda_t}(\mathbf{x})$
end for
return $\mathbf{x}^T$
```

Problem 5 (F-score Revisited: 7+8 = 15 marks)**.** Recall that the F-score of a binary classifier $f : \mathcal{X} \to \{-1,+1\}$ is the harmonic mean of its precision $P(f) = \mathbb{P}\left[Y = 1 \,|\, f(X) = 1\right]$ and recall $R(f) = \mathbb{P}\left[f(X) = 1 \,|\, Y = 1\right]$ i.e. $F(f) := \frac{2P(f)R(f)}{P(f)+R(f)}$. All probabilities are with respect to the underlying data distribution $\mathcal{D}$. Given a set S of n samples $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$ from the data distribution $\mathcal{D}$, design and analyze an algorithm to estimate the F-score of a given classifier f that does not depend on S. Give bounds for your estimate of the form $\left|\hat{F}(f) - F(f)\right| \leq \epsilon$ with probability at least $1 - \delta$. Denote $\mathbb{P}\left[Y = 1\right] = \pi > 0$ and $\mathbb{P}\left[f(X) = 1\right] = \mu > 0$ but we dont know π or μ.

Solution: Define $T = \mathbb{P}\left[Y = 1 \wedge f(X) = 1\right]$. Then we have $P(f) = \frac{T}{\mu}$ and $R(f) = \frac{T}{\pi}$ so that $F(f) = \frac{2T}{\pi+\mu}$. Define the following quantities

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{y_i = 1\right\} \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{f(\mathbf{x}_i) = 1\right\} \qquad \hat{T} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{f(\mathbf{x}_i) = 1 \wedge y_i = 1\right\}$$

and define $\hat{F}(f) = 2\hat{T}/\left(\hat{\pi} + \hat{\mu}\right)$. By an application of Chernoff's bound and a union bound, we know that with probability at least $1 - 3\exp(-n\epsilon^2/3)$, we have $\max\left\{\left|\hat{\pi} - \pi\right|, \left|\hat{\mu} - \mu\right|, \left|\hat{T} - T\right|\right\} \leq \epsilon$. Notice we have used an absolute error bound and not a relative error bound i.e. we have error of the form $\left|X - \hat{X}\right| \leq \epsilon$ rather than $\left|X - \hat{X}\right| \leq \epsilon \cdot X$. We need to be more careful if working with relative bounds since it may be the case that $T = 0$ even if $\pi, \mu > 0$ which will create confidence issues. Now, for all $\epsilon < (\pi + \mu)/4$ we have

$$\hat{F}(f) = \frac{2\hat{T}}{\hat{\pi} + \hat{\mu}} \leq \frac{2T + 2\epsilon}{\pi + \mu - 2\epsilon} \leq F(f) + \frac{8\epsilon}{\pi + \mu}$$
$$\hat{F}(f) = \frac{2\hat{T}}{\hat{\pi} + \hat{\mu}} \geq \frac{2T - 2\epsilon}{\pi + \mu + 2\epsilon} \geq F(f) - \frac{4\epsilon}{\pi + \mu}$$

To get the above expressions, notice that $2T \leq \pi + \mu$ since $T \leq \min\left\{\pi, \mu\right\}$ and make some simple manipulations. Setting $\Delta = \frac{8\epsilon}{\pi+\mu}$, we have $\left|\hat{F}(f) - F(f)\right| \leq \Delta$ with prob. at least $1 - 3\exp(-n(\pi+\mu)^2\Delta^2/192)$, for all $\Delta < 1$.

Problem 6 (Pseudoconvex Programming: 7+8 = 15 marks)**.** Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be two functions s.t. f is convex, g is concave, and $0 < m \leq \left\{f(\mathbf{x}), g(\mathbf{x})\right\} \leq M$ for all $\mathbf{x} \in \mathcal{C}$ where $\mathcal{C} \subseteq \mathbb{R}^d$ is convex. Design and analyze an algorithm to solve $\min_{\mathbf{x}\in\mathcal{C}} \frac{f(\mathbf{x})}{g(\mathbf{x})}$. Assume that you have access to a convex optimization oracle over $\mathcal{C}$ that can return the minimizer of any convex function over $\mathcal{C}$ in one step. State clearly any assumptions you are making.

Solution: The sublevel trick tells us that since $g(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{C}$, we have $h(\mathbf{x}) \leq \lambda \Leftrightarrow f(\mathbf{x}) - \lambda \cdot g(\mathbf{x}) \leq 0$. Define $V_\lambda(\mathbf{x}) = f(\mathbf{x}) - \lambda \cdot g(\mathbf{x})$. Notice that $V_\lambda(\mathbf{x})$ is convex for all $\lambda > 0$ and consider Algorithm 2. Note that $V_{\lambda_t}(\mathbf{x}^t) = 0$ by construction so we must have $V_{\lambda_t}(\mathbf{x}^{t+1}) := -e_t \leq 0$ which implies that $f(\mathbf{x}^{t+1}) = \lambda_t \cdot g(\mathbf{x}^{t+1}) - e_t$.

This gives us

$$h(\mathbf{x}^{t+1}) = \frac{f(\mathbf{x}^{t+1})}{g(\mathbf{x}^{t+1})} = \lambda_t - \frac{e_t}{g(\mathbf{x}^{t+1})} \leq \lambda_t - \frac{e_t}{M}$$

However, we also have $V_{\lambda_t}(\mathbf{x}^*) \geq -e_t$ where $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{C}}\ h(\mathbf{x})$ is the optimum of the problem. This implies $h(\mathbf{x}^*) \geq \lambda_t - \frac{e_t}{g(\mathbf{x}^*)} \geq \lambda_t - \frac{e_t}{m}$ in other words $e_t \geq (\lambda_t - h(\mathbf{x}^*)) \cdot m$. This gives us

$$h(\mathbf{x}^{t+1}) \leq \lambda_t - \frac{e_t}{M} \leq \lambda_t - (\lambda_t - h(\mathbf{x}^*)) \cdot \frac{m}{M}$$

which in turn gives us, since $\lambda_t = h(\mathbf{x}^t)$,

$$h(\mathbf{x}^{t+1}) - h(\mathbf{x}^*) = (h(\mathbf{x}^t) - h(\mathbf{x}^*)) \left(1 - \frac{m}{M}\right),$$

i.e. $h(\mathbf{x}^T) \leq h(\mathbf{x}^*) + (h(\mathbf{x}^0) - h(\mathbf{x}^*)) \left(1 - \frac{m}{M}\right)^T$ which ensures a linear rate of convergence to the optimum function value.