Prof. T. Hofmann

Final Exam

22nd August 2016

First and Last name: ______________________

Student ID (Legi) Nr: ______________________

Signature: ______________________

General Remarks

- Please check that you have all 20 pages of this exam.
- There are 120 points, and the exam is 120 minutes. **Don't spend too much time on a single question!** The maximum of points is not required for the best grade!
- Remove all material from your desk which is not permitted by the examination regulations.
- Write your answers directly on the exam sheets. If you need more space, make sure you put your **student-ID**-number on top of each supplementary sheet.
- Immediately inform an assistant in case you are not able to take the exam under regular conditions. Later complaints are not accepted.
- Attempts to cheat/defraud lead to immediate exclusion from the exam and can have judicial consequences.
- Please use a black or blue pen to answer the questions.
- Provide only one solution to each exercise. Cancel invalid solutions clearly.

	Topic	Max. Points	Points Achieved	Visum
1	Dimensionality Reduction	30		
2	Clustering, Mixture Models, NMF	30		
3	Sparse Coding and Robust-PCA	30		
4	Optimization	30		
Total		120		

Grade: ....

1 Dimensionality reduction (30 pts)

1.1 SVD and PCA

Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

Note that in the following we assume that any given data matrix X is centered (i.e. has zero mean).

7 pts □

a) In PCA, one performs eigenvalue decomposition on the data matrix.
[ ] True [ ] False

b) In PCA, discarding only the zero eigenvalues (and therefore the corresponding eigenvectors) guarantees perfect data reconstruction. [ ] True [ ] False

c) Suppose one performs PCA on a set of points X, and then performs PCA again on a rotation Y of the points X (by an orthogonal matrix V). This always gives the same reconstruction error. [ ] True [ ] False

d) The performance of SVD for collaborative filtering is not dependent on the initialization of the unknown / missing ratings. [ ] True [ ] False

e) Using SVD for collaborative filtering guarantees that different users with exactly the same known/filled ratings will receive the same item recommendations.
[ ] True [ ] False

f) *(counts for 2/-2pts).* Using an SVD for collaborative filtering, movies with similar storylines will have similar vector representations after performing SVD for collaborative filtering.
[ ] True [ ] False

1.2 PCA

Part 1

Let $X = A^\top A$ be a $d \times d$ matrix, where A is an $n \times d$ matrix.

a) Show that the eigenvalues and eigenvectors of the matrix X can be computed using only the SVD decomposition of matrix A.

3 pts □

....

....

. . . .

b) Suppose we want to perform PCA on n data points, where each point is d-dimensional. Suppose that $d >> n$ (e.g. data points represent a few high resolution images in which each pixel is considered to represent a dimension).

Show that, in this case, it is more efficient to perform SVD on the data matrix instead of doing eigen-decomposition on the covariance matrix.

2 pts □

. . . .

. . . .

Part 2

Consider you have 3 data points in a two-dimensional space as follows:
$A = (-1, 1)$, $B = (0, 0)$, and $C = (1, -1)$.

a) What is the first principal component (vector) for this data?

3 pts □

. . . .

. . . .

. . . .

b) Suppose you project the data points in a one-dimensional space represented by the first principal component, what would the coordinates of the new datapoints be?

2 pts □

. . . .

. . . .

c) How large is the variance of the projected data?

2.5 pts □

. . . .

. . . .

....

d) How large is the reconstruction error if you represent the projected data back in the original two-dimensional space?

2.5 pts □

....

....

....

1.3 SVD

Consider a symmetric matrix X with distinct eigenvalues. We denote by u a unit eigenvector of X whose corresponding eigenvalue is λ_1 and v is an eigenvector of X whose corresponding eigenvalue is λ_2. Let $Y = X + uu^\top$.

a) Prove that Y is also symmetric.

2 pts □

....

....

....

b) Show that u is also an eigenvector of Y and find its corresponding eigenvalue.

3 pts □

....

....

....

c) Show that v is also an eigenvector of Y and find its corresponding eigenvalue.

3 pts □

....

. . . .

. . . .

2 Mixture Models, NMF, Word Embedding and Neural Network (30 pts)

2.1 Mixture Model and Expectation Maximization

Suppose we have a set of N data points in d dimensions, $\mathbf{X} := (x_1, \ldots, x_N)$. We want to model this data using a mixture of K Gaussian distributions $\mathcal{N}(x|\mu_k, \Sigma_k)$. The Gaussian mixture model for one data point is thus defined as,

$$p_\theta(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k),$$

where $\theta = (\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ are the model parameters.

a) Briefly explain the differences between K-means and Gaussian mixture model.

3 pts □

. . . .

. . . .

. . . .

b) Show that $p_\theta(x)$ has mean $\langle x \rangle = \sum_{k=1}^{K} \pi_k \mu_k$

2 pts □

. . . .

. . . .

. . . .

c) Write down the expression for the log-likelihood $\mathcal{L}(\mathbf{X}, \theta)$ of the entire dataset $\mathbf{X}$.

2 pts □

. . . .

....

d) The Expectation Maximization algorithm maximizes a lower bound on the log-likelihood. Write down the derivation of this lower bound on $\mathcal{L}(\mathbf{X}, \theta)$ (assuming the posterior distribution to be $q_{k,n}$).

3 pts □

....

....

....

e) Derive the update rule for Σ_k when maximizing the above lower bound.

Reminder: $\mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$.

3 pts □

....

....

....

....

....

2.2 Topic Model and Nonnegative Matrix Factorization

Which of the following claims are true/false? *(2 points per correct answer, non-negative total points in any case)*

10 pts □

a) In probabilistic LSA, the occurrence of word w in document d is represented as

$$p(w|d) = \sum_{z=1}^{K} p(w|d,z)p(z|d) = \sum_{z=1}^{K} p(w|z)p(z|d)$$

where topics are identified with integers $z \in \{1, \cdots, K\}$.

[ ] True [ ] False

b) In probabilistic LSA, assume the data has been summarized into co-occurrence counts $X = (x_{ij})$ (number of occurrences of w_j in document d_i), then the log-likelihood is given by

$$\sum_{i,j} x_{ij} \log \sum_{z=1}^{K} p(w_j|z)p(z|d_i).$$

[ ] True [ ] False

c) Semi-NMF — when additionally requiring all columns of $\mathbf{Z}$ to sum up to one — is equivalent to *soft* k-means clustering.

[ ] True [ ] False

d) A non-negative matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ of $rank(\mathbf{X}) = K < \min\{D, N\}$ can always be factorized as $\mathbf{X} = \mathbf{UZ}$ with non-negative $\mathbf{U} \in \mathbb{R}^{D \times K}$ and $\mathbf{Z} \in \mathbb{R}^{K \times N}$.

[ ] True [ ] False

e) In the NMF Algorithm, the update for $\mathbf{Z}$ is given as

$$z_{kn} \leftarrow z_{kn} \frac{\left(\mathbf{U}^{\top}\mathbf{U}\mathbf{Z}\right)_{kn}}{\left(\mathbf{U}^{\top}\mathbf{X}\right)_{kn}}$$

[ ] True [ ] False

2.3 Word Embedding and Neural Network

a) Given the data summarized in co-occurrence matrix $\mathbf{N} = (n_{ij})$, the GloVe (Global Vectors) objective is the weighted least squares fit of log-counts,

$$\mathcal{H}(\theta; \mathbf{N}) = \sum_{i,j} f(n_{ij})(\log n_{ij} - \langle \mathbf{x}_i, \mathbf{y}_j \rangle)^2,$$

where the bias terms are ignored for simplicity. This objective is generally non-convex, the stochastic gradient descent (SGD) works well in practice. In one iteration, SGD first of all sample (i, j) such that $n_{ij} > 0$ uniformly at random, then perform a "cheap" update using this sample. Let the stepsize be η, write down the update rule for $\mathbf{x}_i$ and $\mathbf{y}_j$.

3 pts □

. . . .

. . . .

. . . .

. . . .

b) Write down two typical activation functions for neural network (both name and formula).

2 pts □

. . . .

. . . .

c) Describe two of the key components/ideas for a convolutional neural network.
Reminder: The formula for a convolution for a 2D image x with 7x7 filter is:

$$F_{n,m}(\mathbf{x};\mathbf{w}) = \sigma(b + \sum_{k=-3}^{3}\sum_{l=-3}^{3} w_{k,l} \cdot x_{n+k,m+l})$$

where (n, m) is the center of receptive field, $\mathbf{w}$ is the weights.

2 pts □

. . . .

. . . .

. . . .

. . . .

3 Sparse Coding and Robust PCA (30 pts)

3.1 Fourier and wavelet transform

Part 1

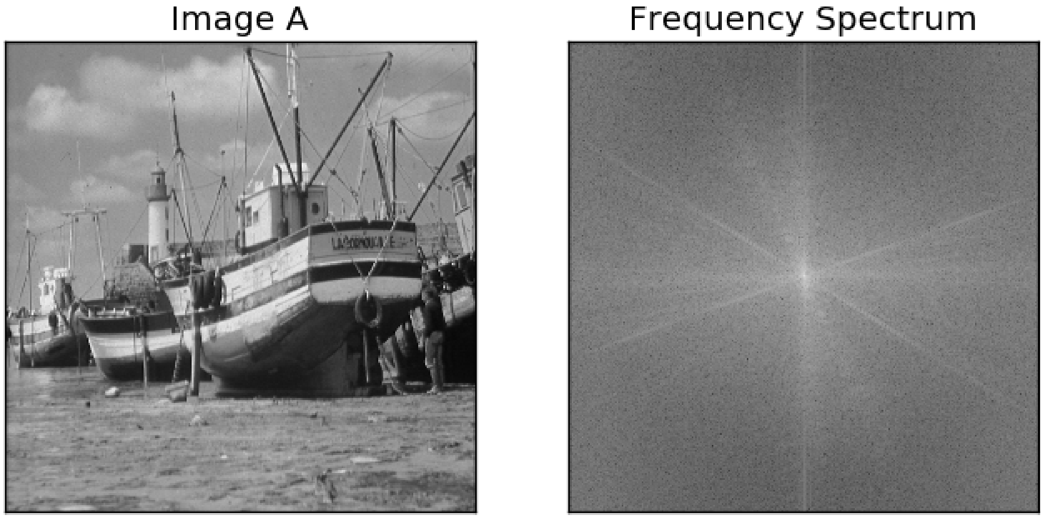
a) Suppose we have two bases, one being orthonormal and the other one being non-orthonormal. What is the advantage of the orthonormal basis for decoding?

3 pts □

....

....

b) We are given the image A with its frequency spectrum shown in the figure below.



Recall that the center of the image corresponds to the magnitude of the low frequencies, while white regions further away from the center correspond to high frequencies.

Denoising with the Fast Fourier Transform (FFT): Here, we are given a noisy version of the original image A. The pseudocode below summarizes the steps to denoise A.

(a) Take the FFT of the image

(b) Retain low [ ] high [ ] frequencies (select one)

(c) ?

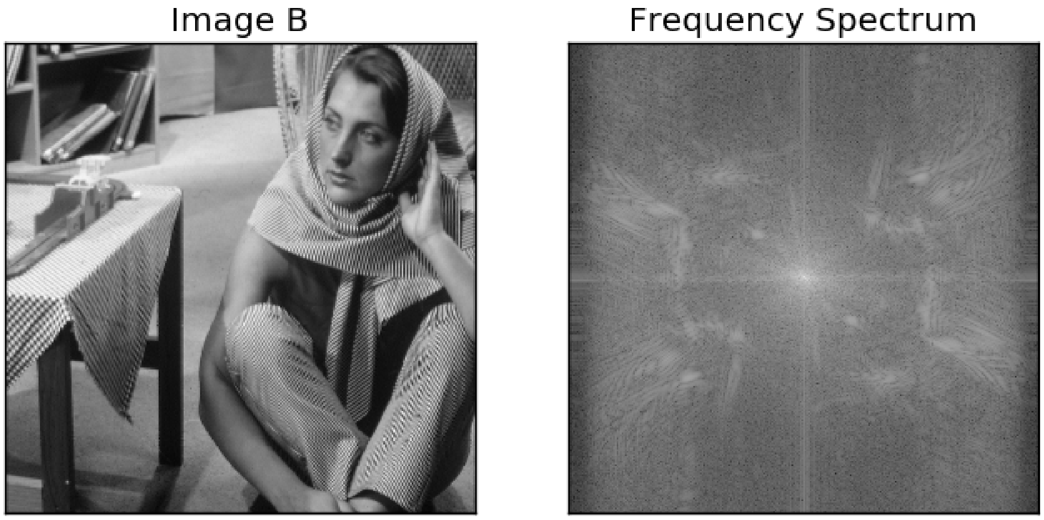
Select one option in line (b) and write down step (c) below.

3 pts □

....

....

c) Now we are given an image B with its frequency spectrum shown in the figure below.



Unlike the frequency spectrum of image A, the spectrum of B has more high-frequency components. Name one reason that could explain this difference.

3 pts □

....

....

d) What are the advantages of using a wavelet basis instead of Fourier to encode images?

2 pts □

....

....

e) Check all the correct statements (0.5 point per correct answer, -0.5 point per incorrect answer, non-negative total points in any case).

If $U \in \mathbb{R}^{D \times D}$ is an orthogonal matrix and $x, y \in \mathbb{R}^D$ then:

2 pts □

[ ] $\operatorname{rank}(U) = D$

[ ] $\det(U) = 0$
[ ] $\|U^\top x\|_2^2 = \|x\|_2^2$
[ ] $U^\top x + U^\top y = x + y$

Part 2

a) Suppose you have the following basis in $\mathbb{R}^2$

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Calculate the coherence of B. Add a normalized vector so that the coherence will be minimum.

2 pts □

. . . .

. . . .

b) Consider the optimization problem:

$$z^* = \arg\min_z \|z\|_0$$

$$\text{s.t. } x = Uz,$$

where $z \in R^n$, $U \in R^{m \times n}$, and $x \in R^m$, $n > m$.
Explain how does this formulation differ from directly solving for $x = Uz$?

2 pts □

. . . .

. . . .

c) Suggest two possible strategies for solving the above problem.

3 pts □

. . . .

. . . .

d) Suggest two applications for the above optimization problem.

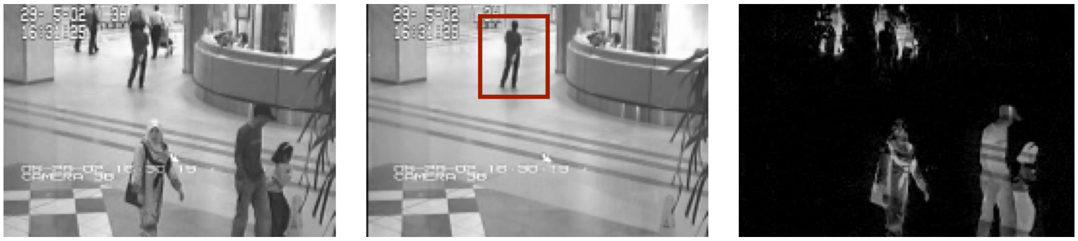
3 pts □

....

....

3.2 Robust PCA

a) We use robust PCA to extract the foreground and background of a video. The left-most image shows one frame extracted from the video with its corresponding background and foreground in the middle and right image respectively.



We see that each person in the video is considered as foreground, except the one in the rectangle. Can you explain why this is the case?

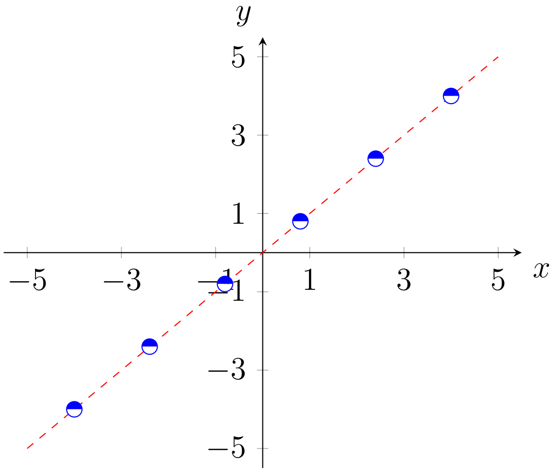
3 pts □

....

....

....

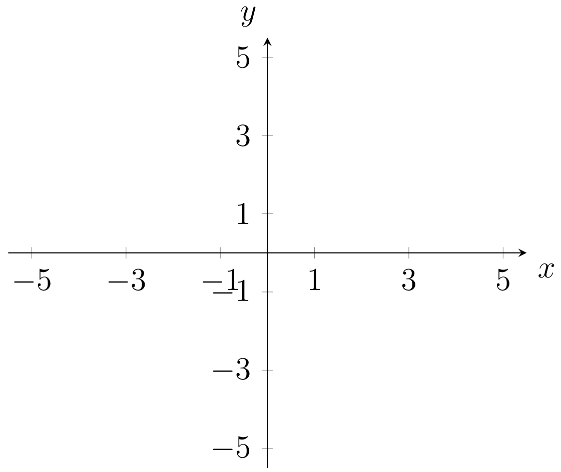
b) We are given the following set of points shown in the figure below. The red line is the principal component of this set, denoted u_1.



Suggest a corrupted version of these points such that PCA can still estimate the principal component u_1. Draw the points on the figure below.

2 pts □

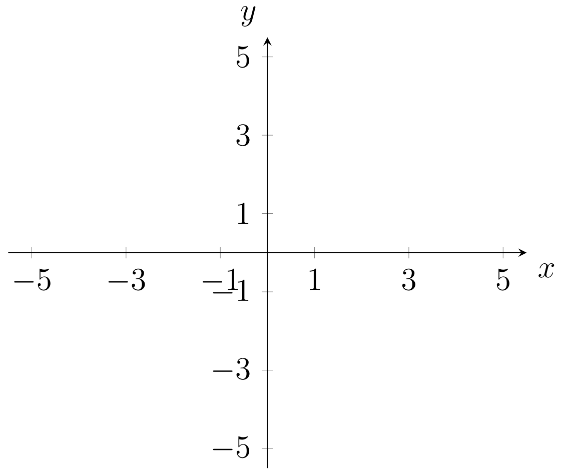
. . . .



Suggest a corrupted version of these points such that Robust PCA estimates the principal component more reliably than PCA. Draw the points on the figure below.

2 pts □

. . . .



4 Optimization (30 pts)

4.1 Convex Optimization

Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

Notation: Consider a vector $\mathbf{x} = \{x_1 \dots x_n\} \in R^n$.

6 pts □

a) Consider two arbitrary vectors $\mathbf{a} = \{a_1 \dots a_n\} \in R^n$ and $\mathbf{b} = \{b_1 \dots b_n\} \in R^n$. Is the following set convex? $\{\mathbf{x} \mid a_i \leq x_i \leq b_i\}$.

[ ] True [ ] False

b) Is the following set convex? $\{\mathbf{x} \mid \sum_{i=1}^{n} x_i^2 = 1\}$

[ ] True [ ] False

c) *(counts for 2/-2pts).* Is $rank(A)$ convex?

[ ] True [ ] False

d) *(counts for 2/-2pts).* Is $\|.\|_0$ convex?

[ ] True [ ] False

4.2 Lagrange Duality

Consider the optimization problem given as $\min \mathbf{c}^\top \mathbf{x}$, $\mathbf{x} \in \mathbb{R}^N$ under the constraints

$$A\mathbf{x} = \mathbf{b}$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u},$$

where $A \in \mathbb{R}^{D \times N}$, $\mathbf{b} \in \mathbb{R}^D$, $\mathbf{l}, \mathbf{u} \in \mathbb{R}^N$.

a) Write down the Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ corresponding to this optimization problem.

2 pts □

....

....

b) Derive the dual function $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ as a function of $\mathbf{x}$.

2 pts □

....

....

c) Find an explicit expression for $\mathbf{x}$ *and* $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$.

3 pts □

....

....

....

4.3 Optimization methods

Part 1

Consider the function $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^2 + \mathbf{y}^2$.

a) We use the shorthand notation $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. What is the minimum point $\mathbf{z}^* = \arg\min_{\mathbf{z}} f(\mathbf{z})$?

2 pts □

....

....

b) Calculate the gradient $\nabla f(\mathbf{z})$.

2 pts □

....

....

c) Apply one step of gradient descent starting at $\mathbf{z} = (2, 2)$ with a learning rate $\eta = \frac{1}{2}$. How many steps are required to get to the minimum $\mathbf{z}^*$?

2 pts □

....

....

d) What happens when $\eta = 1$ and $\eta = \frac{1}{4}$? Make 3 iterations and describe their behaviour.

2 pts □

....

....

e) What condition on the learning rate do we require to guarantee convergence if the evaluation of the gradient $\nabla f(\mathbf{z})$ is inexact (e.g. as in Stochastic Gradient Descent)?

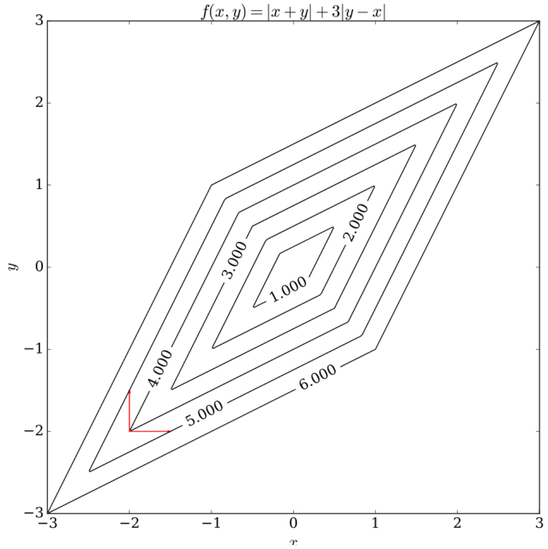
2 pts □

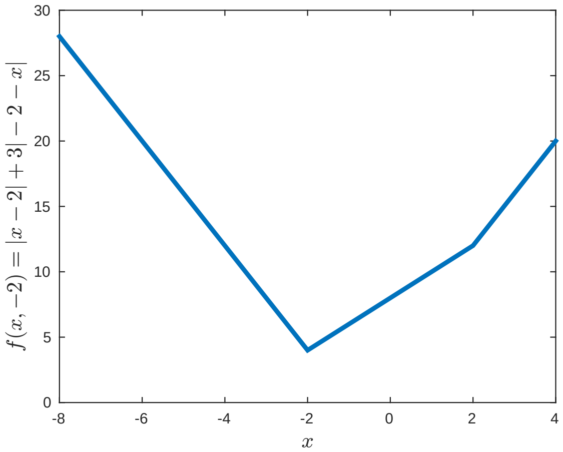
....

....

Part 2

Consider the non-differentiable objective function $f(x, y) = |x+y| + 3|y-x|$ where $x, y \in \mathbb{R}$. A contour plot of $f(x, y)$ as well as a plot of f for a fixed $y = -2$ is shown below.





a) Explain why coordinate descent might fail to optimize this objective. (Hint: start coordinate descent from $(-2, -2)$).

3 pts □

....

....

....

b) Let $A \in R^{n\times d}, \mathbf{y} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^n$. We consider the two functions

$$f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|^2 + \lambda\|\mathbf{x}\|_2$$

and

$$g(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|^2 + \lambda\|\mathbf{x}\|_1.$$

Write down the update of one step of coordinate descent to solve $\min_{\mathbf{x}} f(\mathbf{x})$.

2 pts □

....

....

....

....

Write down the update of one step of coordinate descent to solve $\min_{\mathbf{x}} g(\mathbf{x})$.

2 pts □

....

....

....

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet