Optimization and Algorithms
November 7, 2024

Write your name:	______________________________
Write your student number:	______________________________

Write your answers (A, B, C, D, E, or F) to problems 1 to 3 in this box

Your answer to problem 1: __________

Your answer to problem 2: __________

Your answer to problem 3: __________

Write your answer to problem 4 in pages 5 and 6

Write your answer to problem 5 in pages 7 and 8

Write your answer to problem 6 in pages 9 and 10

Exam

1. *Taking a single photo of an animal. (3 points)* Let $\widehat{p}_1, \ldots, \widehat{p}_T$ represent the predicted trajectory that a given animal is expected to follow from time instant $t = 1$ to time instant $t = T$, where $\widehat{p}_t \in \mathbf{R}^3$ denotes the predicted position of the animal at time t.

We control a drone whose dynamics are given by

$$x_{t+1} = Ax_t + Bu_t, \quad t = 0, 1, \ldots, T-1,$$

where $u_0, u_1, \ldots, u_{T-1}$ represents the control signal (to be designed). The state $x_t \in \mathbf{R}^d$ is partitioned as

$$x_t = \begin{bmatrix} p_t \\ q_t \end{bmatrix},$$

where $p_t \in \mathbf{R}^3$ represents the position of the drone at time t, and $q_t \in \mathbf{R}^{d-3}$ represents other state variables (whose physical meanings are not important in this problem). We know the initial state $x_0 \in \mathbf{R}^d$ and the matrices A and B.

We wish to design the control signal $u_0, \ldots, u_{T-1}$ so that the drone gets as close as possible to the animal at some point of the predicted trajectory (but not necessarily at all points) to take a single photo. In other words, our goal is to design a control signal that guides the drone to approach the animal as closely as possible at some point along its predicted trajectory; this allows for capturing a single photograph of the animal, without requiring the drone to maintain close proximity throughout the entire predicted trajectory.

Let I_3 be the 3×3 identity matrix and define

$$E = \begin{bmatrix} I_3 & 0 \end{bmatrix} \in \mathbf{R}^{3 \times d}.$$

Note that Ex_t gives the position p_t of the drone at time t. Also, as usual, let $\|v\|_2$ denote the Euclidean norm of the vector v (that is, $\|v\|_2 = (v^T v)^{1/2}$).

Consider the optimization problem

$$\begin{array}{ll} \underset{u_0,\ldots,u_{T-1},x_1,\ldots,x_T}{\text{minimize}} & f(u_0,\ldots,u_{T-1},x_1,\ldots,x_T) \\ \text{subject to} & x_{t+1} = Ax_t + Bu_t, \quad t = 0,1,\ldots,T-1, \end{array} \tag{1}$$

with the following choices for the objective function:

(A)

$$f(u_0,\ldots,u_{T-1},x_1,\ldots,x_T) = \min\left\{\|Ex_t - \widehat{p}_s\|_2 \;:\; s,t = 1,\ldots,T\right\}$$

(B)

$$f(u_0,\ldots,u_{T-1},x_1,\ldots,x_T) = \sum_{t=1}^{T} \|Ex_t - \widehat{p}_t\|_2$$

(C)

$$f(u_0,\ldots,u_{T-1},x_1,\ldots,x_T) = \max\left\{\|Ex_t - \widehat{p}_t\|_2 \;:\; t = 1,\ldots,T\right\}$$

(D)

$$f(u_0,\ldots,u_{T-1},x_1,\ldots,x_T) = \min\left\{\|Ex_t - \widehat{p}_t\|_2 \;:\; t = 1,\ldots,T\right\}$$

(E)

$$f(u_0,\ldots,u_{T-1},x_1,\ldots,x_T) = \max\left\{\|Ex_t - \widehat{p}_s\|_2 \;:\; s,t = 1,\ldots,T\right\}$$

(F)

$$f(u_0,\ldots,u_{T-1},x_1,\ldots,x_T) = \sum_{s,t=1}^{T} \|Ex_t - \widehat{p}_s\|_2$$

One of these cost functions is most suitable for the given context. Which one?

Write your answer (A, B, C, D, E, or F) in the box at the top of page 1

2. *Weighted least-squares. (3 points)* Weighted least-squares with regularization is a problem of the form

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|_D^2 + \rho\,\|x\|_2^2\ , \tag{2}$$

where the variable to optimize is $x \in \mathbf{R}^n$.

The matrix $A \in \mathbf{R}^{m \times n}$, the vector $b \in \mathbf{R}^m$, the constant $\rho > 0$, and the diagonal matrix

$$D = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_m \end{bmatrix} \in \mathbf{R}^{m \times m}$$

(with $d_i > 0$ for $1 \le i \le m$) are given. For a vector

$$u = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \in \mathbf{R}^m,$$

the notation $\|u\|_D^2$ means $u^T D u$ (equivalently, $d_1 u_1^2 + \cdots + d_m u_m^2$). Assume that A has linearly independent columns and let I_n denote the $n \times n$ identity matrix.

One of the following vectors is the solution of (2):

(A) $(A^T D A + \rho I_n)^{-1} A^T D b$
(B) $(A^T D^2 A + \rho I_n) A^T D^2 b$
(C) $(A^T D^2 A + \rho I_n)^{-1} A^T D^2 b$
(D) $(A^T A)^{-1} A^T b$
(E) $(A^T D A + \rho I_n) A^T D b$
(F) $(A^T D^{-1} A + \rho I_n)^{-1} A^T D^{-1} b$

Which one?

Write your answer (A, B, C, D, E, or F) in the box at the top of page 1

3. *Newton method with step-size fixed at* 1 *(2 points)* Consider the quadratic function $f : \mathbf{R}^2 \to \mathbf{R}$ given by

$$f(x) = (x - b)^T A (x - b) + c,$$

where

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad c = 5.$$

Suppose that the Newton method with the step-size fixed at 1 is used to minimize f.

If

$$x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

is the initial iterate, then the next iterate x_1 is one of the following points:

(A) $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$

(B) $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$

(C) $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$

(D) $\begin{bmatrix} -2 \\ -2 \end{bmatrix}$

(E) $\begin{bmatrix} 0 \\ -1 \end{bmatrix}$

(F) $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$

Which one?

Write your answer (A, B, C, D, E, or F) in the box at the top of page 1

Optimization and Algorithms

November 7, 2024

Write your name:	____________________
Write your student number:	____________________

4. *Linear regression with output priors. (4 points)* Let the time-series $x_1, x_2, \ldots, x_T$ be given, where $x_t \in \mathbf{R}^d$ for $1 \leq t \leq T$. We want to learn a parameter $w \in \mathbf{R}^d$ that maps linearly this input time-series to another output time-series, denoted by $y_1, \ldots, y_T$, where $y_t \in \mathbf{R}$ is given by $y_t = w^T x_t$ for $1 \leq t \leq T$.

The output time-series should have a given mean μ (that is, $(\sum_{t=1}^T y_t)/T = \mu$) and its empirical variance should be, at most, a given constant σ^2 (that is, $(\sum_{t=1}^T (y_t - \mu)^2)/T \leq \sigma^2$). We have only noisy measurements of the output sequence, which we denote by $\widehat{y}_1, \ldots, \widehat{y}_T$, where $\widehat{y}_t \in \mathbf{R}$ for $1 \leq t \leq T$.

In this context, fitting the parameter w can be carried out by solving the optimization problem

$$\begin{array}{ll} \underset{w}{\text{minimize}} & \sum_{t=1}^T \left(w^T x_t - \widehat{y}_t\right)^2, \\ \text{subject to} & \frac{\sum_{t=1}^T w^T x_t}{T} = \mu \\ & \frac{\sum_{t=1}^T \left(w^T x_t - \mu\right)^2}{T} \leq \sigma^2. \end{array} \tag{3}$$

Show that (3) is a convex optimization problem.

Optimization and Algorithms

November 7, 2024

Write your name:	________________________
Write your student number:	________________________

5. *Disaggregating a load. (4 points)* Load disaggregation means guessing the energy consumed over time by each appliance in a household, given measurements of their aggregate consumption.

When the energy consumption of the appliances are modelled as uncorrelated Gaussian random variables, load disaggregation leads to the following optimization problem:

$$\begin{array}{ll} \underset{x_1,\ldots,x_N}{\text{minimize}} & \frac{1}{2}\sum_{n=1}^{N} (x_n - \mu_n)^T C_n^{-1} (x_n - \mu_n) \\ \text{subject to} & x_1 + \cdots + x_N = y, \end{array} \tag{4}$$

where the variables to optimize are $x_1, \ldots, x_N$, with $x_n \in \mathbf{R}^d$ for $1 \leq n \leq N$.

In (4), the vectors $\mu_n \in \mathbf{R}^d$ and the $d \times d$ positive-definite matrices C_n are given. The vector $y \in \mathbf{R}^d$ is also given. (The vector $x_n \in \mathbf{R}^d$ represents the energy consumed by appliance n at d time instants; μ_n and C_n are the mean and the covariance matrix of the underlying prior Gaussian model for appliance n; and $y \in \mathbf{R}^d$ represents the total consumption measured at d time instants.)

Give the solution of (4) in closed-form.

(Hint: if you have trouble solving this problem, try to solve it for the simpler case $d = 1$. In this simple case, each $d \times d$ matrix C_n becomes just a positive number. You can still get partial score for this simpler case.)

Optimization and Algorithms

November 7, 2024

Write your name:	______________________
Write your student number:	______________________

6. *Personalized learning. (4 points)* Personalized learning involves problems of the form

$$\underset{x_1,\ldots,x_N}{\text{minimize}} \quad \underbrace{\sum_{n=1}^{N} f_n(x_n) + \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} w_{nm} \left\| x_n - x_m \right\|_2^2}_{f(x_1,\ldots,x_N)}, \tag{5}$$

where the variables to optimize are $x_1, \ldots, x_N$, with $x_n \in \mathbf{R}^d$ for $1 \leq n \leq N$.

In (5), each $f_n : \mathbf{R}^d \to \mathbf{R}$ is a given loss function. The constants $w_{nm} \geq 0$ are also given. (Each f_n can be interpreted as a loss function localized at person n, who wishes to minimize f_n by choosing the local model parameter x_n; the inclusion of the soft penalty terms $w_{nm} \left\| x_n - x_m \right\|_2^2$ encourages the optimal model parameters to be similar across all the persons.)

Suppose that $f_n : \mathbf{R}^d \to \mathbf{R}$ is a strongly convex function, for $1 \leq n \leq N$.

Show carefully that the objective function

$$f : \underbrace{\mathbf{R}^d \times \cdots \times \mathbf{R}^d}_{N \text{ factors}} \to \mathbf{R}$$

of problem (5) is a strongly convex function.

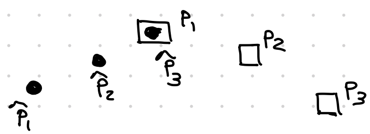
Optimization and Algorithms, IST

Nov 2024, João Xavier

Solution of the Exam

Problem 1 **Answer: D**

<u>Details</u> (a) In this option, the minimum can be achieved without both entities being at the same location at the same time. Example:



(b) This option tries to minimize the averange distance between p_t and $\hat{p}_t$

(c) This option tries to minimize the distance between p_t and $\hat{p}_t$ throughout the whole time horizon

(e) This option tries to minimize the maximal distance between the two trajectories, disregarding the time stamps

(f) This option tries to minimize the averge distance between the two trajectories, ignoring the time stamps.

Problem 2 **Answer: A**

Details The function $f(x) = \|Ax-b\|_D^2 + \rho\|x\|_2^2$ is equivalent to the quadratic function

$$f(x) = (Ax-b)^T D(Ax-b) + \rho x^T x,$$

whose gradient is $\nabla f(x) = 2A^T D(Ax-b) + 2\rho x = 2\left(A^T DA + \rho I\right)x - 2A^T Db$.

Setting the gradient to zero gives

$$\nabla f(x) = 0 \iff x = \left(A^T DA + \rho I\right)^{-1} A^T Db.$$

Because f is convex, this stationary point is a global minimizer.

Problem 3 **Answer: B**

Details The Newton method with the step size fixed at 1 is the method

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k).$$

For $f(x) = (x-b)^T A(x-b) + c$, we have $\nabla f(x) = 2A(x-b)$ and $\nabla^2 f(x) = 2A$.

Thus,

$$x_1 = x_0 - (2A)^{-1}(2A(x_0-b)) = b.$$

In our case, we were given that $b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

Problem 4

A problem of the form

$$\min_{w} \quad f(w)$$
$$\text{s.t.} \quad h(w) = 0$$
$$\qquad g(w) \le 0$$

is convex if f and g are convex and h is affine.

In our case, $f(w) = \sum_{t=1}^{T} (w^T x_t - \hat{y}_t)^2$, $h(w) = \frac{1}{T}\left(\sum_{t=1}^{T} w^T x_t\right) - \mu$, and

$g(w) = \frac{1}{T}\sum_{t=1}^{T} (w^T x_t - \mu)^2 - \sigma^2$.

$\boxed{\text{Analysis of } f}$

f can be expressed as $f(w) = \| X^T w - \hat{y} \|_2^2$, where

$$X = [x_1 \cdots x_T] : d \times T \text{ and } \hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_T \end{bmatrix} \in \mathbb{R}^T.$$

Thus, f is the composition $f = p \circ q$, where $q: \mathbb{R}^d \to \mathbb{R}^T$, $q(w) = X^T w - \hat{y}$, and $p(u) = \|u\|_2^2$.

Note that q is an affine map; also, $p(u) = u^T I u$ is convex (because it is a quadratic and $I \succeq 0$). We conclude that f is convex.

$\boxed{\text{Analysis of } h}$

h can be written as $h(w) = w^T \bar{x} - \mu$, where $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$. Thus, h is an affine function.

$\boxed{\text{Analysis of } g}$

Write g as $g = g_0 + \frac{1}{T} g_1 + \cdots + \frac{1}{T} g_T$, where $g_0(w) = -\mu$ and $g_t(w) = (w^T x_t - \mu)^2$.

If $g_0, g_1, \dots, g_T$ are convex, then g is convex.

The function g_0, being affine (actually, a constant), is convex.

Consider now g_t for $1 \le t \le T$. We have $g_t = q_t \circ p_t$, where $p_t(w) = w^T x_t - \mu$ and $q_t(u) = u^2$. Because p_t is affine and q_t is convex, g_t is convex.

Problem 5

Start by rewriting the problem

$$\min_{x_1, \dots, x_N} \quad \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu_n)^T C_n^{-1} (x_n - \mu_n)$$

$$\text{s.t.} \quad x_1 + \dots + x_N = y$$

in the more compact form

$$\min_{x} \quad \underbrace{1/2 \, (x-\mu)^T C^{-1} (x-\mu)}_{f(x)},$$

$$\text{s.t.} \quad Ex = y,$$

where $x = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}$, $C = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_N \end{bmatrix}$, and $E = [I \cdots I]$.

The KKT system for this problem is

$$\begin{cases} \nabla f(x) = E^T \lambda \\ Ex = y \end{cases} \iff \begin{cases} C^{-1}(x-\mu) = E^T \lambda \\ Ex = y. \end{cases}$$

To solve the system, the first equation gives $x = \mu + CE^T\lambda$. Plugging this expression in the second equation generates the equation

$$E\mu + ECE^T\lambda = y,$$

whose solution is

$$\lambda = (ECE^T)^{-1}(y - E\mu) = (C_1 + \cdots + C_N)^{-1}(y - (\mu_1 + \cdots + \mu_N)).$$

Finally, using this expression in $x = \mu + CE^T\lambda$ gives

$$x = \mu + CE^T\lambda$$

$$\Leftrightarrow \quad \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix} + \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_N \end{bmatrix} \begin{bmatrix} I \\ \vdots \\ I \end{bmatrix} \lambda$$

$$\Leftrightarrow \quad \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \mu_1 + C_1\lambda \\ \vdots \\ \mu_N + C_N\lambda \end{bmatrix}$$

$$\Leftrightarrow \quad \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \mu_1 + C_1(C_1 + \cdots + C_N)^{-1}(y - (\mu_1 + \cdots + \mu_N)) \\ \vdots \\ \mu_N + C_N(C_1 + \cdots + C_N)^{-1}(y - (\mu_1 + \cdots + \mu_N)) \end{bmatrix}.$$

This solution of the KKT system is a global minimizer of our problem because the problem is convex.

Problem 6

Write the function $f: \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$,

$$f(x_1, \ldots, x_N) = \sum_{n=1}^{N} f_n(x_n) + \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} w_{nm} \|x_n - x_m\|_2^2,$$

as $f = g + h$, where $g: \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$, $g(x_1, \ldots, x_N) = \sum_{n=1}^{N} f_n(x_n)$, and $h: \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$,
$h(x_1, \ldots, x_N) = \sum_{n=1}^{N} \sum_{m=n+1}^{N} w_{nm} \|x_n - x_m\|_2^2$.

If we show that g is strongly convex and h is convex, then f is strongly convex.

Analysis of g

By definition, g is strongly convex if there exists $m>0$ such that

$$g((1-\alpha)x+\alpha y) \le (1-\alpha)g(x) + \alpha g(y) - m\frac{\alpha(1-\alpha)}{2}\|x-y\|_2^2$$

holds for all $x=(x_1,\dots,x_N)\in\mathbb{R}^d\times\cdots\times\mathbb{R}^d$, $y=(y_1,\dots,y_N)\in\mathbb{R}^d\times\cdots\times\mathbb{R}^d$ and $0\le\alpha\le 1$.

Consider $n\in\{1,2,\dots,N\}$.

Because $f_n:\mathbb{R}^d\to\mathbb{R}$ is strongly convex, there exists $m_n>0$ such that

$$f_n((1-\alpha)x_n+\alpha y_n) \le (1-\alpha)f_n(x_n) + \alpha f_n(y_n) - m_n\frac{\alpha(1-\alpha)}{2}\|x_n-y_n\|_2^2$$

holds for all $x_n\in\mathbb{R}^d$, $y_n\in\mathbb{R}^d$, and $0\le\alpha\le 1$.

Now, let $m=\min\{m_1,\dots,m_N\}$. Note that $m>0$. Also,

$$f_n((1-\alpha)x_n+\alpha y_n) \le (1-\alpha)f_n(x_n) + \alpha f_n(y_n) - m\frac{\alpha(1-\alpha)}{2}\|x_n-y_n\|_2^2$$

holds for all $x_n\in\mathbb{R}^d$, $y_n\in\mathbb{R}^d$, and $0\le\alpha\le 1$.

Summing the N inequalities above (for $n=1,2,\dots,N$) gives

$$\underbrace{\sum_{n=1}^{N} f_n((1-\alpha)x_n+\alpha y_n)}_{g((1-\alpha)x+\alpha y)} \le (1-\alpha)\underbrace{\left(\sum_{n=1}^{N} f_n(x_n)\right)}_{g(x)} + \alpha\underbrace{\left(\sum_{n=1}^{N} f_n(y_n)\right)}_{g(y)} - m\frac{\alpha(1-\alpha)}{2}\underbrace{\left(\sum_{n=1}^{N}\|x_n-y_n\|_2^2\right)}_{\|x-y\|_2^2},$$

which shows that g is strongly convex.

Analysis of h

The function $h:\mathbb{R}^d\times\cdots\times\mathbb{R}^d\to\mathbb{R}$ is a linear combination, with nonnegative weights w_{nm}, of functions of the form $h_{nm}:\mathbb{R}^d\times\cdots\times\mathbb{R}^d\to\mathbb{R}$, $h_{nm}(x_1,\dots,x_N)=\|x_n-x_m\|_2^2$.

If we show that h_{nm} is convex, then h is convex.

Write h_{nm} as $h_{nm}(x)=\|Ax\|_2^2$, where $A=[0\cdots 0\ I_d\ 0\cdots 0\ -I_d\ 0\cdots 0]: d\times Nd$.

$$\begin{array}{ccc} & \uparrow & \uparrow \\ & \text{block } n & \text{block } m\end{array}$$

The function h_{nm} is the composition $h_{nm}=q_{nm}\circ p_{nm}$, where $p_{nm}:\mathbb{R}^d\times\cdots\mathbb{R}^d\to\mathbb{R}$, $p_{nm}(x)=Ax$, and $q_{nm} : \mathbb{R}^d \to \mathbb{R}$, $q_{nm}(a) = \|a\|_2^2$. Because p_{nm} is an affine map, and q_{nm} is convex, the function h_{nm} is convex.