Exam: Optimization for Machine Learning

M2 IASD/MASH

Thursday, January 6, 2022

This is an open book exam, meaning you can consult any written or printed material. Electronic devices are prohibited. You must justify all of your answers. If you cannot solve a question, do not hesitate to admit its result in order to solve the next ones.

Part 1 (MASH+IASD)

The following exercises deal with the first part of the course (basics of optimization) and should be addressed by both the IASD and MASH students.

Ex. 1 — For $f : \mathbb{R} \to \mathbb{R}$, we define

$$g : (u, v) \in \mathbb{R}^2 \mapsto g(u, v) \coloneqq f(u\, v).$$

We want to study the gradient descent on $\mathbb{R}^2$ of g

$$(u_{k+1}, v_{k+1}) = (u_k, v_k) - \tau \nabla g(u_k, v_k).$$

To simplify this analysis, we consider instead its continuous time counterpart, $t \mapsto (u(t), v(t))$, which satisfies the following differential equation

$$\left(\frac{\mathrm{d}u}{\mathrm{d}t}(t), \frac{\mathrm{d}v}{\mathrm{d}t}(t)\right) = -\nabla g(u, v) \quad \text{and} \quad u(0) = u_0,\ v(0) = v_0.$$

We denote $x(t) \coloneqq u(t)v(t) \in \mathbb{R}$.

1. Compute $\nabla g(u, v)$ as a function of $f'(x)$ and (u, v) where $x = uv$.
2. Show that
$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = -H(u(t), v(t))\, f'(x(t))$$
where $H : \mathbb{R}^2 \to \mathbb{R}$ is a positive function to be determined.
3. Show that for all $t \geq 0$, $u(t)^2 - v(t)^2 = u_0^2 - v_0^2$. In the following, we denote $A \coloneqq u_0^2 - v_0^2$.
4. Given that $u(t)^2 - v(t)^2 = A$ and $u(t)v(t) = x(t)$, find expressions of $u(t)^2$ and $v(t)^2$ as functions of $(x(t), A)$. Deduce an expression of $H(u(t), v(t))$ as a function of $x(t)$ and A. Conclude that $x(t)$ satisfies an ODE of the form
$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = -G(x(t), A)\, f'(x(t)) \quad \text{and} \quad x(0) = x_0 \coloneqq u_0 v_0$$
where G is a positive function to be determined.

5. When A is small, how would you compare qualitatively the behaviour of this evolution with respect to the usual gradient flow $\tilde{x}(t)$ solution of

$$\frac{\mathrm{d}\tilde{x}}{\mathrm{d}t}(t) = -f'(\tilde{x}(t)) \quad \text{and} \quad \tilde{x}(0) = x_0.$$

Ex. 2 — In this exercise, we consider a general, unconstrained optimization problem of the form

$$\text{minimize}_{w\in\mathbb{R}^d} f(w), \tag{1}$$

where $f \in \mathcal{C}^1$. Rather than assuming that the gradient is Lipschitz continuous (as in the lectures), we will assume a more general property called *Hölder continuity*. More precisely, for any $\nu \in (0, 1]$ and $L > 0$, we write $f \in \mathcal{C}_L^{1,\nu}$ if $f \in \mathcal{C}^1$ and

$$\forall (v, w) \in (\mathbb{R}^d)^2, \qquad \|\nabla f(v) - \nabla f(w)\| \leq L\|v - w\|^\nu. \tag{2}$$

We say that ∇f is (L, ν)-Hölder continuous.
For any $f \in \mathcal{C}_L^{1,\nu}$, we can show that

$$\forall (v, w) \in (\mathbb{R}^d)^2, \qquad f(v) \leq f(w) + \nabla f(w)^T(v - w) + \frac{L}{1+\nu}\|v - w\|^{1+\nu}. \tag{3}$$

Moreover, since $f \in \mathcal{C}^1$, we can apply gradient descent to problem (1): the kth iteration of this method is

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k). \tag{4}$$

We first study the case $\nu = 1$ (i.e. $f \in \mathcal{C}_L^{1,1}$), corresponding to the gradient being L-Lipschitz continuous.

1. Recalling that $f \in \mathcal{C}_L^{1,1}$ for this question, justify the choice $\alpha_k = \frac{1}{L}$ for every k.
2. Under this assumption, give the convergence rate result for gradient descent on problem (1), assuming that the function f is
   a) nonconvex;
   b) convex.
3. Assuming that f is convex, name a method with a better convergence rate than gradient descent, and give the associated rate.

We now consider the more general setting $\nu \in (0, 1]$ (i.e. $f \in \mathcal{C}_L^{1,\nu}$) and f nonconvex.

4. Based on (3) and your answer to Question 1, justify the choice

$$\alpha_k = \left[\frac{\|\nabla f(w_k)\|^{1-\nu}}{L}\right]^{\frac{1}{\nu}} \tag{5}$$

   for the stepsize used in the iteration (4).
5. Using the sequence $\{\alpha_k\}_k$ from the previous question, it is possible to show a complexity bound in $\mathcal{O}\left(\epsilon^{-\frac{1+\nu}{\nu}}\right)$ for gradient descent applied to $f \in \mathcal{C}_L^{1,\nu}$ nonconvex and bounded below. Compare this bound with the results seen in class for the case $\nu = 1$.
6. In addition to $f \in \mathcal{C}_L^{1,\nu}$, suppose that f is nonconvex and twice continuously differentiable. Suppose also that gradient descent converges to $w^* \in \mathbb{R}^d$.

a) What can we say about $\nabla f(w^*)$? Is w^* a local minimum?
b) Using that $f \in \mathcal{C}^2$, what additional guarantee seen in class can be provided about w^*?

Ex. 3 — In this exercise, we consider a dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^{d_x}$ and $y_i \in \mathbb{R}^{d_y}$: our goal is to learn a model that predicts y_i out of x_i. To this end, we consider a two-layer linear network

$$\begin{array}{rccl} h : \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_y} \\ x & \longmapsto & W_2(W_1 x + b_1) + b_2, \end{array} \tag{6}$$

where $W_1 \in \mathbb{R}^{d_h \times d_x}$, $W_2 \in \mathbb{R}^{d_y \times d_h}$, $b_1 \in \mathbb{R}^{d_h}$ and $b_2 \in \mathbb{R}^{d_y}$.

$$\text{minimize}_{w \in \mathbb{R}^d} \; f(w) := \frac{1}{n} \sum_{i=1}^n f_i(w), \quad \text{where} \quad f_i(w) := \frac{1}{2} \left\| h(x_i; w) - y_i \right\|^2. \tag{7}$$

It can be shown that every component function f_i is continuously differentiable.

1. Write down an iteration of batch stochastic gradient for problem (7) (with an arbitrary stepsize).
2. How do gradient descent and stochastic gradient arise as special cases of the batch stochastic gradient from the previous question?
3. Give one argument in favor of choosing a batch size larger than 1, and one argument against it.
4. We consider that our main computational cost lies in accessing an example (x_i, y_i) from the dataset. Using that metric, what is the cost of an iteration of gradient descent, and that of an iteration of batch stochastic gradient descent?

The problem (7) is nonconvex: a standard convergence rate analysis for stochastic gradient descent leads to a result of the form

$$\mathbb{E}\left[\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \right] \; \leq \; \mathcal{O}\left(\tfrac{1}{K^{1/4}}\right) \tag{8}$$

for any $K \geq 1$. Meanwhile, the standard analysis for gradient descent on nonconvex problems gives

$$\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \; \leq \; \mathcal{O}\left(\tfrac{1}{K^{1/2}}\right) \tag{9}$$

for any $K \geq 1$.

5. Compare the guarantees (8) and (9) for
   a) a fixed number of iterations;
   b) a fixed number of epochs.
6. What observation from the previous question can be used as theoretical motivation for using stochastic gradient?

We finally discuss extensions of the basic (batch) stochastic gradient framework.

7. Recall that gradient aggregation methods such as SAGA combine a stochastic gradient estimate with a full gradient approximation maintained over several iterations. What is the intended goal of such an approach?
8. Describe two other ways of incorporating information from the previous iterations, that are both used in one of the methods described during the lectures. What is this method?

Ex. 4 — Let $a_1, \dots, a_N$ be pairwise distinct points in $\mathbb{R}^p$. We want to compute a *geometric median* of $a_1, \dots, a_N$, that is, we want to find a solution of

$$\min_{x \in \mathbb{R}^p} f(x) \quad \text{where} \quad f(x) \stackrel{\text{def}}{=} \sum_{i=1}^{N} \|x - a_i\| \tag{10}$$

and $\|\cdot\|$ is the usual Euclidean norm (Notice that there is NO SQUARE on the norms!). That problem is also known as the Weber problem.

1. a) Prove that there exists a solution to (10).
   b) In general, is the solution unique?
   Hint: *You may consider the case $p = 2$, $N = 2$, $a_1 = (0, 0)$, $a_2 = (1, 0)$.*
2. In this question, we study the auxiliary function $g \colon \mathbb{R}^p \to \mathbb{R}$ defined by $g(x) = \|x\|$.
   a) Prove that g is convex, and that its dual conjugate is given by

$$\forall s \in \mathbb{R}^p, \quad g^*(s) = \begin{cases} 0 & \text{if } \|s\| \leq 1, \\ +\infty & \text{otherwise.} \end{cases} \tag{11}$$

   b) Recall the Fenchel inequality and its equality case. Deduce a necessary and sufficient condition for $s \in \partial g(x)$.
   Hint: *You may distinguish between the cases $x = 0$ and $x \neq 0$.*
3. a) Write the optimality conditions for (10) and prove that they are equivalent to the following.
   - A point $x \notin \{a_1, \dots, a_N\}$ is a solution if and only if

$$0 = \sum_{i=1}^{N} \frac{x - a_i}{\|x - a_i\|}.$$

   - A point $x = a_{i_0}$, with $1 \leq i_0 \leq N$, is a solution if and only if

$$\left\| \sum_{i \neq i_0} \frac{x - a_i}{\|x - a_i\|} \right\| \leq 1.$$

   b) If $x \notin \{a_1, \dots, a_N\}$ and x is a solution, deduce that x must belong to the convex hull of $a_1, \dots, a_N$.
4. Find a solution in the following cases, for $p = 2$:
   a) $a_1 = (0, 0)$, $a_2 = (1, 0)$, $a_3 = (1, 1)$, $a_4 = (0, 1)$ (you may simply justify by a drawing)
   b) $a_1 = (0, 0)$, $a_2 = (1, 0)$, $a_3 = (-\frac{1}{2}, \frac{\sqrt{3}}{2})$. ***Hint:*** *In that case, please justify carefully that a_1 is a solution.*
5. We want to prove that the solution to (10) is unique if the points $a_1, \dots, a_N$ are not all aligned.
   a) Prove that the solution set to (10) is convex. Deduce that if there are at least two minimizers, then there is a minimizer $x^\star \notin \{a_1, \dots, a_N\}$.

b) Prove that the function f is twice differentiable at $x^\star$ and that, if $a_1, \dots, a_N$ are not all aligned, its Hessian is positive definite, that is $\nabla^2 f(x^\star) \succ 0$.
Hint: *You may use/prove the fact that the auxiliary function satisfies*

$$\forall h \in \mathbb{R}^p, \quad \langle \nabla^2 g(x)h, h\rangle = \frac{1}{\|x\|}\left(\|h\|^2 - \left(\left\langle \frac{x}{\|x\|}, h\right\rangle\right)^2\right) = \frac{\left\|P_{\{x\}^\perp} h\right\|^2}{\|x\|},$$

where $P_{\{x\}^\perp} h$ is the orthogonal projection of h onto the orthogonal space to x.

c) Conclude.

Part 2 (IASD only)

The following exercises deal with the second part of the course and should only be addressed by IASD students.

Ex. 5 — For $u \in \mathbb{R}^k$, let $f(u) \coloneqq \log(\sum_{i=1}^k \exp(u_i))$ (log-sum-exp).

1. Show that
$$\nabla f(u) = \frac{\exp(u)}{\sum_{i=1}^k \exp(u_i)} = \frac{\exp(u)}{\exp(f(u))}, \tag{12}$$
(softmax). Note that $\exp(u)$ means that exp is applied component-wise to u.
2. For $v \in \mathbb{R}^k$, recall that the conjugate is defined by
$$f^*(v) \coloneqq \max_{u \in \mathbb{R}^k} \langle u, v \rangle - f(u). \tag{13}$$
Show that
$$f^*(v) = \langle v, \log v \rangle = \sum_{i=1}^k v_i \log(v_i) \tag{14}$$
with domain $\mathrm{dom}(f^*) = \{v \in \mathbb{R}^k \colon v \geq 0, \sum_{i=1}^k v_i = 1\}$ (the probability simplex). We assume $0 \log 0 = 0$. Hint: set the gradient w.r.t. u of $\langle u, v \rangle - f(u)$ to zero, in order to relate v with u, then substitute for v in $\langle u, v \rangle - f(u)$.
3. For $\theta_i, y_i \in \mathbb{R}^k$, let $L(\theta_i, y_i) \coloneqq f(\theta_i) - \langle \theta_i, y_i \rangle$ (logistic loss). Reusing f^*, write the expression of the conjugate in the first argument
$$L^*(-\alpha_i, y_i) = \max_{\theta_i \in \mathbb{R}^k} \langle -\alpha_i, \theta_i \rangle - L(\theta_i, y_i) \tag{15}$$
4. For $\theta \in \mathbb{R}^{n \times k}$ a matrix gathering the θ_i, let $F(\theta) \coloneqq \sum_{i=1}^n L(\theta_i, y_i)$ be the sum of losses. Let $G(W) \coloneqq \frac{\lambda}{2}\|W\|^2$. Consider the objective
$$\min_{W \in \mathbb{R}^{d \times k}} F(XW) + G(W), \tag{16}$$
where $X \in \mathbb{R}^{n \times d}$ is a feature matrix.
Recall that the Fenchel dual is
$$\max_{\alpha \in \mathbb{R}^{n \times k}} -F^*(-\alpha) - G^*(X^\top \alpha). \tag{17}$$
Using the above, write down the dual expression in the specific case when L is the logistic loss. Make sure to indicate what the constraints on α are.

Ex. 6 — Consider the following problem
$$p^* = \min_{x,z \in \mathbb{R}} x^2 - 3xz + z^2 + 2x, \quad \text{subject to} \quad x + z = 1 \tag{18}$$

1. Write the Lagrangian of the problem above. Prove that p^* is always larger than

$$\min_{x,z\in\mathbb{R}} x^2 - 3xz + z^2 - 2z + 2$$

   (hint: use weak duality and bound from below the maximum with a specific value for the dual variable)
2. Prove that the problem in (18) admits a saddle point and that strong duality holds.
3. Write the algorithm to solve (18) by dual gradient ascent.

Answer (Ex. 1) — 1. $\nabla f(u,v) = (vf'(uv), uf'(uv))^\top = (v,u)^\top f'(x)$.

2. One has
$$\dot{x}(t) = v(t)\dot{u}(t) + u(t)\dot{v}(t) = -(u(t)^2 + v(t)^2)f'(x(t)),$$
so that $H(u,v) = u^2 + v^2$.

3. Denote $w(t) \coloneqq u(t)^2 - v(t)^2$, so that
$$\dot{w}(t) = \dot{u}(t)u(t) - \dot{v}(t)v(t) = v(t)f'(x(t))u(t) - u(t)f'(x(t))v(t) = 0.$$

4. Omitting the dependency on t, one has
$$uv = x \quad \text{and} \quad u^2 - v^2 = A.$$
Denoting $U \coloneqq u^2$, one has the equation
$$U - \frac{x^2}{U} = A \quad \Rightarrow \quad U^2 - AU - x^2 = 0 \quad \Rightarrow \quad U = \frac{A + \sqrt{A^2 + 4x^2}}{2},$$
hence
$$H(u,v) = u^2 + v^2 = 2u^2 - A = 2\frac{A + \sqrt{A^2 + 4x^2}}{2} - A = \sqrt{A^2 + 4x^2}.$$

5. Since G is positive, both $\tilde{x}$ and x are gradient flow and $f(\tilde{x}(t))$ and $f(x(t))$ are decaying. If A is small, $G(x,A) \approx 2|x(t)|$ so that the gradient descent is evolving more slowly if x is small.

Answer (Ex. 2) — 1. By choosing $\alpha = \frac{1}{L}$, the gradient iteration (4) becomes
$$w_{k+1} = w_k - \frac{1}{L}\nabla f(w_k).$$
Using this relation in (3) with $\nu = 1$, we obtain:
$$\begin{aligned} f(w_{k+1}) &\leq f(w_k) + \nabla f(w_k)^T(w_{k+1} - w_k) + \frac{L}{2}\|w_{k+1} - w_k\|^2 \\ &= f(w_k) - \frac{1}{L}\|\nabla f(w_k)\|^2 + \frac{1}{2L}\|\nabla f(w_k)\|^2 \\ &= f(w_k) - \frac{1}{2L}\|\nabla f(w_k)\|^2. \end{aligned}$$
The choice $\alpha_k = \frac{1}{L}$ for every k thus guarantees a decrease in the function value as long as $\|\nabla f(w_k)\| \neq 0$. *Bonus point: The value $\frac{1}{L}$ actually minimizes the right-hand side of the first inequality.*

2. The convergence rate of gradient descent for $\mathcal{C}^{1,1}$ functions is
a) $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ if f is nonconvex, in the sense that for any $K \geq 1$, we have
$$\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

b) $\mathcal{O}(\frac{1}{K})$ if f is convex, in the sense that for any $K \geq 1$, we have

$$f(w_k) - \min_{w \in \mathbb{R}^d} f(w) \leq \mathcal{O}\left(\frac{1}{K}\right).$$

3. Provided f is convex, the accelerated gradient method (also called Nesterov's method) possesses a better convergence rate than gradient descent on convex problems: this rate is of order $\mathcal{O}\left(\frac{1}{K^2}\right)$.
4. We apply (3) for the prescribed choice of α_k:

$$\begin{aligned} f(w_{k+1}) &\leq f(w_k) + \nabla f(w_k)^T(w_{k+1} - w_k) + \frac{L}{1+\nu}\|w_{k+1} - w_k\|^{1+\nu} \\ &= f(w_k) - \left[\frac{\|\nabla f(w_k)\|^{1-\nu}}{L}\right]^{\frac{1}{\nu}} \|\nabla f(w_k)\|^2 + \frac{L}{1+\nu}\left[\frac{\|\nabla f(w_k)\|^{1-\nu}}{L}\right]^{\frac{1+\nu}{\nu}} \|\nabla f(w_k)\|^{1+\nu} \\ &= f(w_k) - \frac{\|\nabla f(w_k)\|^{\frac{1+\nu}{\nu}}}{L^{\frac{1}{\nu}}} + \frac{1}{1+\nu}\frac{\|\nabla f(w_k)\|^{\frac{1+\nu}{\nu}}}{L^{\frac{1}{\nu}}} \\ &= f(w_k) - \frac{\nu}{1+\nu}\frac{\|\nabla f(w_k)\|^{\frac{1+\nu}{\nu}}}{L^{1/\nu}}. \end{aligned}$$

As a result, the chosen value for α_k yields a decrease in the function value as long as $\|\nabla f(w_k)\| \neq 0$: this justifies to use such a value as a stepsize. *Bonus point: As for the case $\nu = 1$, the chosen value actually minimizes the right-hand side of the first inequality.*

5. For $f \in \mathcal{C}_L^{1,1}$ nonconvex and bounded below, the convergence rate of gradient descent is $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$, that translates to a complexity bound in $\mathcal{O}(\epsilon^{-2})$, i.e. the method produces an iterate with gradient norm smaller than ϵ in at most $\mathcal{O}(\epsilon^{-2})$ iterations.
Replacing $\mathcal{C}_L^{1,1}$ by $\mathcal{C}_L^{1,\nu}$, we obtain a complexity in $\mathcal{O}(\epsilon^{-\frac{1+\nu}{\nu}})$, that matches the one seen in class for $\nu = 1$. For $\nu < 1$, we have $\epsilon^{-\frac{1+\nu}{\nu}} > \epsilon^{-2}$ for $\epsilon < 1$, hence the complexity bound is worse than in the case $\nu = 1$.
Alternatively, the answer can be turned into a convergence rate for the comparison.
6. *For this question, we assume $f \in \mathcal{C}_L^{1,\nu}$ and $f \in \mathcal{C}^2$.*
   a) Since gradient descent converges to w^*, this point must be a first-order stationary point, i.e. $\|\nabla f(w^*)\| = 0$. Without any convexity assumptions on f, this point could be a local minimum, a saddle point, or a local maximum.
   b) Since the objective is twice continuously differentiable, we know that for almost any initial point $w_0 \in \mathbb{R}^d$, a limit point of gradient descent will be a second-order stationary point, i.e. w^* must satisfy $\nabla f(w^*) = 0$ and $\nabla^2 f(w^*) \succeq 0$.

Answer (Ex. 3) — 1. The kth iteration of a batch stochastic gradient method is

$$w_{k+1} = w_k - \frac{\alpha_k}{|\mathcal{S}_k|}\sum_{i \in \mathcal{S}_k} \nabla f_i(w_k),$$

where $\alpha_k > 0$ is a stepsize for iteration k, w_k is the current iterate, and $\mathcal{S}_k$ is a random set of indices drawn in $\{1, \ldots, n\}$.

2. If the set $\mathcal{S}_k$ consists in a single element, then the batch stochastic gradient method becomes the classical stochastic gradient algorithm. On the other hand, if $\mathcal{S}_k = \{1, \dots, n\}$ for every k, the iteration corresponds to that of gradient descent.
3. *This answer is more exhaustive than what was asked for in the question.* Using a batch size larger than 1 can help in producing gradient estimates with lower variance, and even lead to faster convergence by considering more samples. In a distributed setting, it can also enable to exploit parallelism. But in a centralized setting, using a batch size remains more expensive in terms of accesses to data points, while being prone to capture redundancies in the data. Finally, a batch size larger than 1 may not necessarily fall into a mini-batch regime and can lead to a slow convergence behavior similar to that of gradient descent.
4. An iteration of gradient descent accesses all n examples of the dataset, yielding a cost of n. Meanwhile, an iteration of stochastic gradient accesses only one example (x_i, y_i) at every iteration: with that metric, it has a cost of 1.
5. *Comparison of* (8) *and* (9)
   a) For a fixed number of iterations K, stochastic gradient guarantees a convergence rate in $\mathcal{O}\left(\frac{1}{K^{1/4}}\right)$ in expectation, while gradient descent's deterministic rate is $\mathcal{O}\left(\frac{1}{K^{1/2}}\right)$. By observing that the sequence $\left\{\frac{1}{K^{1/2}}\right\}_K$ goes to 0 faster than $\left\{\frac{1}{K^{1/4}}\right\}_K$ as $K \to \infty$, we conclude that gradient descent provides a better guarantee (of finding a point with small gradient norm) than stochastic gradient.
   b) We recall that one epoch corresponds to one iteration of gradient descent and n iterations of stochastic gradient. Given a number of epochs N_e, the rate (9) for gradient descent is that after N_e iterations, hence $\mathcal{O}\left(\frac{1}{N_e^{1/2}}\right)$. Meanwhile, the guarantee for stochastic gradient after N_e epochs corresponds to (8) after $n \times N_e$ iterations, which is $\mathcal{O}\left(\frac{1}{n^{1/4} N_e^{1/4}}\right)$. When $n \gg N_e$, the second guarantee is better (in the sense of being smaller), i.e. the average guarantee of stochastic gradient is better than that of gradient descent for the same number of epochs.
6. The second guarantee shows that, for the same cost in terms of accesses on the data (or passes in the case of an epoch), stochastic gradient yields better error guarantees on average for large amounts of data points. This observation can serve as a motivation for applying stochastic gradient techniques in practice.
7. The goal of these methods is to reduce the variance of the stochastic gradient approximations.
8. One possible way to incorporate information from the previous iteration consists in combining the stochastic gradient direction (produced via one component or a batch thereof) with a *momentum term*, obtained from the displacement at the previous iteration. Another possibility consists in applying *diagonal scaling* to the stochastic gradient step, i.e. to use a different stepsize for every coordinate: this scaling procedure can be done according to the coordinates of the past stochastic gradients, thereby incorporating information from the past iterations. The ADAM variant of stochastic gradient incorporates both momentum and diagonal scaling aspects by exploiting information from the previous iterations.

Answer (Ex. 4) — 1. a) The function $f\colon x \mapsto \sum_{i=1}^N \|x - a_i\|$ is continuous on $\mathbb{R}^p$. More-

over, since $\|x - a_i\| \geq \|x\| - \|a_i\|$, we have

$$f(x) \geq N \|x\| - \sum_{i=1}^{N} \|a_i\|,$$

hence $\lim_{\|x\| \to \infty} f(x) = +\infty$ and f is coercive.
As a result, there exists a minimizer.

b) For any $x \in \mathbb{R}^p$, by the triangle inequality,

$$f(x) = \|x - a_1\| + \|x - a_2\| \geq \|a_2 - a_1\|.$$

Moreover, for all $\theta \in [0, 1]$, we set $x_\theta = (1 - \theta)a_1 + \theta a_2$. Then

$$f(x_\theta) = \theta \|a_2 - a_1\| + (1 - \theta) \|a_2 - a_1\| = \|a_2 - a_1\|.$$

Hence every point in the line segment $[a_1, a_2]$ is a solution.

2. a) By the triangle inequality and homogeneity of the norm,

$$\forall x, y \in \mathbb{R}^p, \forall \theta \in [0, 1], \ \|(1 - \theta)x + \theta y\| \leq (1 - \theta) \|x\| + \theta \|y\|$$

hence g is convex. Moreover,

$$g^*(s) = \sup_{x \in \mathbb{R}^p} (\langle x, s \rangle - \|x\|)$$
$$= \begin{cases} 0 & \text{if } \|s\| \leq 1 \text{ (by the Cauchy-Schwarz inequality, with equality for } x = 0), \\ +\infty & \text{otherwise (choose } x = t\frac{s}{\|s\|} \text{ with } t \to +\infty). \end{cases}$$

b) The Fenchel inequality states that for all $x, s \in \mathbb{R}^p$,

$$g(x) + g^*(s) \geq \langle x, s \rangle$$

with equality if and only $s \in \partial g(x)$.
In our case, this amounts to $\|s\| \leq 1$ and $\langle x, s \rangle = \|x\|$. In other words

- if $x = 0$, it simply means that $\|s\| \leq 1$.
- if $x \neq 0$, the equality case in Cauchy-Schwarz inequality yields $s = \frac{x}{\|x\|}$.

3. The characterization of optimality is $0 \in \partial f(x)$. The functions $g_i \colon x \mapsto \|x - a_i\|$ are convex, proper and l.s.c. (they are continuous on $\mathbb{R}^p$) and $\bigcap_{i=1}^N \operatorname{ri}(\operatorname{dom}(g_i) = \bigcap_{i=1}^N \mathbb{R}^p \neq \emptyset$, hence optimality is equivalent to $0 \in \partial g_1(x) + \cdots + \partial g_N(x)$.
Applying the characterization of the previous question (with $\partial g_i(x) = \partial g(x - a_i)$), we get, for $x \notin \{a_1, \ldots, a_N\}$,

$$0 = \sum_{i=1}^{N} \frac{x - a_i}{\|x - a_i\|}.$$

And for $x = a_{i_0}$, we get

$$0 = s_{i_0} + \sum_{i \neq i_0} \frac{x - a_i}{\|x - a_i\|} \quad \text{for some } s_{i_0} \text{ with } \|s_{i_0}\| \leq 1$$

which is equivalent to

$$\left\| \sum_{i \neq i_0} \frac{x - a_i}{\|x - a_i\|} \right\| \leq 1.$$

a) The optimality condition yields

$$x = \frac{1}{\left(\sum_{j=1}^N \frac{1}{\|x-a_j\|}\right)} \left(\sum_{i=1}^N \frac{a_i}{\|x - a_i\|} \right) = \sum_{i=1}^N \theta_i a_i \tag{19}$$

with $\theta_i \stackrel{\text{def}}{=} \frac{1}{\|x-a_i\|} / \left(\sum_{j=1}^N \frac{1}{\|x-a_j\|}\right)$, so that $\theta_i \geq 0$ and $\sum_i \theta_i = 1$. The point x is a convex combination of the a_i's.

4. a) A (the) solution is at $x = (1/2, 1/2)$ since $\sum_{i=1}^4 \frac{x-a_i}{\|x-a_i\|} = 0$.

b) We check that for $x = a_1$, we have

$$\left\| \frac{x - a_2}{\|x - a_2\|} + \frac{x - a_3}{\|x - a_3\|} \right\|^2 = \left\| \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1/2 \\ \frac{\sqrt{3}}{2} \end{pmatrix} \right\|^2 = 1$$

hence a_1 is a solution.

5. a) If $x, y \in \mathbb{R}^p$ are minimizers, then

$$\forall \theta \in [0,1], \quad \min f \leq f((1-\theta)x + \theta y) \leq (1-\theta) f(x) + \theta f(y) = \min f$$

Hence $x_\theta = (1-\theta)x + \theta y$ is a minimizer.
As a result, the solution set is convex, and if there are two minimizers, there is in fact an infinity of minimizers. We may choose one that is not in $\{a_1, \ldots, a_N\}$.

b) g is square root of a strictly positive C^∞ function on $\mathbb{R}^p \backslash \{0\}$. We compute the derivatives of g at $x \neq 0$,

$$\frac{\partial g}{\partial x_i}(x) = \frac{x_i}{\sqrt{x_1^2 + \cdots + x_N^2}}, \quad \frac{\partial^2 g}{\partial x_i \partial x_j}(x) = \frac{\delta_{i,j}}{\sqrt{x_1^2 + \cdots + x_N^2}} - \frac{x_i x_j}{\left(\sqrt{x_1^2 + \cdots + x_N^2}\right)^3}$$

with $\delta_{i,j} = 1$ if $i = j$, 0 otherwise.
As a result,

$$\langle \nabla^2 g(x) h, h \rangle = \frac{1}{\|x\|} \|h\|^2 - \frac{1}{\|x\|} \left(\langle \frac{x}{\|x\|}, h \rangle \right)^2.$$

We deduce that

$$\langle \nabla^2 f(x) h, h \rangle = \sum_{i=1}^N \frac{\left\| P_{\{x-a_i\}^\perp} h \right\|^2}{\|x - a_i\|}. \tag{20}$$

We see that $\nabla^2 f(x)$ is positive semi-definite, and if $\langle \nabla^2 f(x) h, h \rangle = 0$, then $P_{\{x-a_i\}^\perp} h = 0$ for all i, hence $h \in \bigcap_{i=1}^N \mathrm{Vect}(x - a_i)$. Since not all points are aligned, that intersection is $\{0\}$, and $\nabla^2 f(x) \succ 0$.

c) If $x^\star \notin \{a_1, \dots, a_N\}$, then $\nabla f(x^\star) \succ 0$, so $x^\star$ is an isolated minimizer. But we have seen that if there are at least two minimizers, the set of minimizers is convex, hence $x^\star$ cannot be isolated. Contradiction.

Answer (Ex. 6) — 1. The Lagrangian is

$$L(x, y, z) = x^2 - 3xz + z^2 + 2x + y(x + z - 1).$$

By weak duality we have

$$\max_y \min_{x,z} L(x, y, z) \leq \min_{x,z} \max_y L(x, y, z) =: p^*$$

we choose $y = -2$, obtaining

$$\min_{x,z} x^2 - 3xz + z^2 - 2z + 2 = \min_{x,z} L(x, -2, z) \leq \max_y \min_{x,z} L(x, y, z).$$

2. Since the constraint is affine and the function to be minimized is convex and differentiable, then we can apply the KKT conditions. We have

$$\begin{aligned}
\nabla_x L(x, y, z) &:= 2x - 3z + 2 + y \\
\nabla_y L(x, y, z) &:= x + z + 1 \\
\nabla_z L(x, y, z) &:= -3x + 2z + y.
\end{aligned}$$

By solving the linear system

$$\begin{aligned}
2x - 3z + 2 + y &= 0 \\
x + z + 1 &= 0 \\
-3x + 2z + y &= 0,
\end{aligned}$$

we obtain the saddle point $x^* = -7/10, y^* = -3/2, z^* = -3/10$. The existence of a saddle point guarantees strong duality.

3.
```
x =0, y = 0, z = 0
for t=1 to T
    #in parallel on machine 1, load y and
    x = argmin_x (x^2 -3xz + 2x + yx)
    #on machine 2, load y and
    z = argmin_z (-3xz + z^2 + yz)
    #collect x, z from machine 1 and 2 and
    y = y + gamma(t) * (x+z-1)
```