Convex Optimization Exam

14/06/2024
3 hours – documents allowed
Answers in English or French

The exam is made of 2 independent exercises, and a problem. If necessary, you can admit the results of previous questions. When using the recalls, cite them. "Classifying" an optimization problem consists in precising in which of the category presented in chapter 5 it falls (LP, QP, QCQP, SOCP, SDP, unconstrained or not, differentiable or not, continuous or not, convex or not).

Total number of points 24

Some useful recalls

i) The Fenchel transform of a function f is defined as $f^\star : x^\star \mapsto \sup_{x\in\mathbb{R}^n} \langle x^\star, x\rangle - f(x)$.

ii) If f is a proper, convex, lsc fonction of $\mathbb{R}^n$ to $\mathbb{R}\cup\{+\infty\}$, then for all $x, x^\star \in \mathbb{R}^n$, the following assertion are equivalent:

a) $x^\star \in \partial f(x)$

b) $x \in \partial f^\star(x^\star)$

c) $f(x) + f^\star(x^\star) = \langle x^\star, x\rangle$

d) $x \in \arg\max_{y\in\mathbb{R}^n} \langle x^\star, y\rangle - f(y)$

e) $x^\star \in \arg\max_{y^\star\in\mathbb{R}^n} \langle y^\star, x\rangle - f^\star(y^\star)$

iii) A function Ψ is ρ-convex if, for all x, y, we have $\Psi(y) \geq \Psi(x) + \langle \nabla\Psi(x), y - x\rangle + \frac{\rho}{2}\|y - x\|^2$.

Exercice 1: Breakfast *2 points*

(a) (½ point) Give 2 general ideas that speed-up the convergence of (variants of) the stochastic gradient algorithm.

(b) (1½ points) Consider the following problem. Is it qualified? Give the KKT conditions. Are they enough to solve the problem? Solve the problem without solving the KKT conditions.

$$\begin{aligned} \min_{(x,y)\in\mathbb{R}^2} \quad & xy \\ \text{s.t. } & y \geq x^2 \\ & y \leq x^3 \end{aligned}$$

Solution:

a) Batching, adaptive step-size, momentum...

b) The gradient of the constraint functions are $\nabla g_1(x,y) = (2x, -1)$ and $\nabla g_2(x,y) = (-3x^2, 1)$. If $(x,y) \neq 0$, we have $x, y \geq 1$ and the gradient are positively linearly independent (as $x \geq 1$), thus the constraints are qualified, except at $(0,0)$.

The Lagrangian reads:

$$\mathcal{L}(x, y, \mu_1, \mu_2) = xy + \mu_1(x^2 - y) + \mu_2(y - x^3)$$

The KKT conditions are: (0.5 points)

$$\begin{aligned}
y + 2\mu_1 x - 3\mu_2 x^2 &= 0 && \nabla_x \mathcal{L} = 0 \\
x - \mu_1 + \mu_2 &= 0 && \nabla_y \mathcal{L} = 0 \\
y &\geq x^2 && \text{primal feasability} \\
y &\leq x^3 && \\
\mu_1, \mu_2 &\geq 0 && \text{dual feasability} \\
\mu_1(x^2 - y) &= 0 && \text{complementarity} \\
\mu_2(y - x^3) &= 0 &&
\end{aligned}$$

To solve the problem you have to find the solution to KKT conditions, compare their values and the value at $(0,0)$.

As $y \geq x^2$ is positive, so is $x \geq y$. Thus the cost is positive and $x = y = 0$ is the optimal solution.

Exercice 2: How to save cloud costs? *5 points*

You are working for a company that conceive car design. You have been recruited, as an optimization expert, in the team that optimize a given design. The design is described by a vector $x \in \mathbb{R}^{250}$ of parameters. We are looking for parameters x that minimize the drag coefficient $D(x)$ of the car, under some mass constraints $M(x) \leq \bar{M}$, and budget constraint $B(x) \leq \bar{B}$. The function D, M and B are "smooth" with respect to the parameter x, but their evaluation is numerically expensive (about 1 minute per evaluation per function). There is also a method to evaluate the sensitivity of each function with respect to the parameters which require about (10 minutes per evaluation per function).

All these evaluations are running in the cloud, which is costly. Your boss want to reduce this cost, and count on your knowledge to do so.

Currently, the optimization is done as follows: take the initial design x^0, choose parameter α and β, compute D, M and B at x^0, and the sensitivities s_D, s_M, and s_B. Then update the design as $x^{k+1} = x^k - \frac{1}{\tau}(s_D + \alpha s_M + \beta s_B)$, and repeat the process until the sensitivities are small. If the final design is not satisfying the constraints, the process is restarted by adjusting α and β by hand.

(a) (½ point) Classify the type of problem that is solved by the optimization process. Describe the current methodological approach with the language of optimization.

(b) (1 point) Currently, the parameter τ is hard-coded based on some physical constant. Your boss want to increase it aggressively to reduce the number of iterations. Is it the right approach? What is the risk of doing so? And if you decrease it?

(c) (1 point) He has heard that there might be a way to automatically adjust τ during the optimization process. Give the pseudo-code of such a method you would like to try, and an estimation of the computing time if you try 5 τ before accepting one.

(d) (1 point) Give two other ideas you could try to reduce the cost of the optimization process (for given α and β).

(e) (1 point) Another pain-point is adjusting the parameters α and β by hand. Identify a method to adjust them automatically, and give its pseudo-code (assuming that the method solving the problem for given α and β is given).

(f) (½ point) With the new ecological malus, the constraint on the mass is now a hard constraint (satisfied by the initial design). Suggest a new approach which ensure that the mass constraint is satisfied at each iteration.

Solution:

a) The problem is a differentiable optimization problem with constraints.

$$\begin{aligned} \min_{x \in \mathbb{R}^{250}} \quad & D(x) \\ \text{s.t.} \quad & M(x) \leq \bar{M} \\ & B(x) \leq \bar{B} \end{aligned}$$

The current method consists in dualizing the constraint, and solving the dualized problem with a gradient algorithm with fixed step size $1/\tau$:

$$\min_{x \in \mathbb{R}^{250}} D(x) + \alpha M(x) + \beta B(x)$$

b) Increasing τ will decrease the step-size, so render the approach slower. Decreasing it will increase the step-size, and might lead to divergence.

c) A possible method is to use a backtracking search algorithm to adjust τ at each iteration, starting from a large step (small τ). Each evaluation of a given τ require 1 minute per function, so the

computing time is about 15 minutes. In particular, note that you do not require to recompute any gradient to adjust τ.

Data: Current design $x^{(k)}$, α, β, search direction d, initial step τ_0
Result: Admissible step $1/\tau$
Define $f = D + \alpha M + \beta B$;
$\tau = \tau_0$;
while $f(x^{(k)} + \frac{1}{\tau}d) > f(x^{(k)}) + \frac{1}{2\tau}\big(\nabla f(x^{(k)})^\top d\big)$ **do**
  $\tau = 2\tau$;

Algorithm 1: backtracking linear search for τ

d) We can use another direction for the update, like a *BFGS/Quasi-Newton update*, or a *conjugate gradient* update, even better if *using a well known package* instead of writing it by hand. Newton's method cannot be used as the Hessian is not available. We can also try to *precondition* the problem to have a better convergence.

e) We can use Uzawa's algorithm to adjust the parameters α and β.

Data: Initial design $x^{(0)}$, dual step γ
Result: Optimal design $x^{(k)}$
$\alpha^{(0)} = \beta^{(0)} = 0$;
while $k < 1$ *or* $M(x^{(k)}) > \bar{M}$ *or* $B(x^{(k)}) > \bar{B}$ **do**
  Define $f^{(k)} = D + \alpha^{(k)} M + \beta^{(k)} B$;
  Solve $\min_x f^{(k)}$, starting from $x^{(k)}$, returning $x^{(k+1)}$;
  $\alpha^{(k+1)} := \Big(\alpha^{(k)} + \gamma(M(x^{(k+1)}) - \bar{M})\Big)^+$;
  $\beta^{(k+1)} := \Big(\beta^{(k)} + \gamma(B(x^{(k+1)}) - \bar{B})\Big)^+$;

Algorithm 2: Uzawa algorithm

f) We can use an inner penalization method, where the mass constraint is added to the objective function with $-\ln(\bar{M} - M(x))$ as a penalty.

Exercice 3: Mirror descent and Bregman divergence *17 points*

In this exercise we are going to study a new class of optimization algorithm, called mirror descent, which leverage the Bregman divergence.

The exercise is divided in several parts. Each part of the exercise can be tackled independently of the others, but the results of the previous questions can be used if clearly cited. The part are roughly in increasing order of difficulty, but the first questions of each should be easy to answer.

We say that a function $\Psi : \mathbb{R}^n \to \bar{\mathbb{R}}$ is a *mirror map* if:

i. Ψ is proper, strictly convex, lower semicontinous, and differentiable on the interior of its domain;
ii. the gradient of Ψ diverges on the boundary of its domain (i.e. $\lim_{x \to bd(\operatorname{dom}\Psi)} \|\nabla\Psi(x)\| = +\infty$).

Throughout the exercise we consider that Ψ is a mirror map.

We also consider a proper, convex, lower semicontinuous objective function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, differentiable on its domain, and $x^{(0)} \in \operatorname{dom}(f)$.

(a) We define the associated *Bregman Divergence* associated to Ψ as

$$D_\Psi(x, x_0) = \Psi(x) - \Big(\Psi(x_0) + \langle \nabla\Psi(x_0), x - x_0\rangle\Big). \tag{1}$$

i. (½ point) Give a geometric interpretation of the Bregman divergence (a graph can be useful).

ii. (1 point) Show that, for all $x, x_0 \in \mathrm{dom}(\Psi)$, $D_\Psi(x, x_0) \geq 0$, and that $D_\Psi(x, x_0) = 0$ if and only if $x = x_0$.

iii. (½ point) We admit that $\Psi^\star$ is differentiable on the interior of its domain. Show that, for $x \in \mathrm{dom}(\Psi)$, and $x^\star \in \mathrm{dom}(\Psi^\star)$, the following relations hold:

$$\nabla\Psi^\star(\nabla\Psi(x)) = x \qquad \text{and} \qquad \nabla\Psi(\nabla\Psi^\star(x^\star)) = x^\star$$

Solution:

i. **Geometric interpretation:** The Bregman divergence is the difference between the value of the function at x and the value of the first order approximation of the function at x_0.

ii. **Positivity:** The function Ψ is convex, so $\Psi(x) \geq \Psi(x_0) + \langle \nabla\Psi(x_0), x - x_0\rangle$. The Bregman divergence is the difference between the two sides of this inequality.

Strict convexity: The function Ψ is strictly convex, so the function $x \mapsto \Psi(x) - \langle \nabla\Psi(x_0), x - x_0\rangle$ is strictly convex. Which ensure that $D_\Psi(x, x_0) = 0$ if and only if $x = x_0$.

iii. Direct consequences of recall ii).

(b) We now discuss the mirror descent algorithm, in the unconstrained case.

i. (½ point) We consider the following algorithm, for $\gamma > 0$,

$$x^{(k+1)} = \arg\min_{x \in \mathbb{R}^n} f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)}\rangle + \frac{1}{2\gamma}\|x - \ x^{(k)}\|_2^2.$$

Is the algorithm well-defined? Recognize the algorithm.

ii. (1 point) In its simplest form, the mirror descent algorithm, for unconstrained minimization, is defined as

$$x^{(k+1)} = \arg\min_{x \in \mathbb{R}^n} f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)}\rangle + \frac{1}{\gamma} D_\Psi(x, x^{(k)}).$$

Justify that an iteration of the mirror descent algorithm can be written as

$$x^{(k+1)} = \nabla\Psi^\star\Big(\nabla\Psi(x^{(k)}) - \gamma\nabla f(x^{(k)})\Big).$$

iii. (½ point) Show that $\Psi = \frac{1}{2}\|\cdot\|_2^2$ is a mirror map. What is the mirror descent algorithm associated with this choice of Ψ?

iv. (2 points) We now consider $\Psi = \sum x_i \ln x_i$, for $x \in \mathbb{R}^n_+$, with the convention $0\ln(0) = 0$, and $+\infty$ otherwise. Show that Ψ is a mirror map. Give an explicit formula for the mirror descent algorithm associated with this choice of Ψ.

Solution:

i. The algorithm is well-defined, as the objective function is strongly convex. By looking at the optimality condition, we see that the algorithm is the gradient algorithm with fixed step γ.

ii. By writing the optimality condition we get

$$\begin{aligned} 0 &= \nabla f(x^{(k+1)}) + \frac{1}{\gamma}\nabla_x D_\Psi(\cdot, x^{(k)})(x^{(k+1)}) \\ 0 &= \nabla f(x^{(k+1)}) + \frac{1}{\gamma}\Big(\nabla\Psi(x^{(k+1)}) - \nabla\Psi(x^{(k)})\Big) \\ \nabla\Psi(x^{(k+1)}) &= \nabla\Psi(x^{(k)}) - \gamma\nabla f(x^{(k)}) \end{aligned}$$

Question 1.e) gives the result.

iii. The function $\Psi = \frac{1}{2}\|\cdot\|^2$ is differentiable on the interior of its domain, and its gradient is the identity. The mirror descent algorithm associated with this choice of Ψ is the gradient descent algorithm.

iv. The function $\Psi = \sum x_i \ln x_i$ is differentiable on the interior of its domain, its gradient is $\nabla\Psi(x) = \ln(x) + 1$, and its Hessian is $\nabla^2\Psi(x) = \mathrm{diag}(1/x) > 0$. Thus Ψ is strictly convex, and differentiable on the interior of its domain. (1 point)

Starting from the formula in b), we have (1 points)

$$\begin{aligned} \nabla\Psi(x^{(k+1)}) &= \nabla\Psi(x^{(k)}) - \gamma\nabla f(x^{(k)}) \\ \ln(x_i^{(k+1)}) + 1 &= \ln(x_i^{(k)}) + 1 - \gamma\nabla f(x^{(k)})_i && \forall i \in [n] \\ x_i^{(k+1)} &= x_i^{(k)} e^{-\gamma\nabla f(x^{(k)})_i} && \forall i \in [n] \end{aligned}$$

also known as the exponentiated gradient algorithm.

(c) To address the constrained case, we introduce the notion of Bregman projection. Let $C \subset \mathrm{dom}(\Psi)$ be a closed convex compact set, and $x_0 \in \mathrm{dom}(\Psi)$. The Bregman projection of x_0 on C is defined as

$$\Pi_C(x_0) = \arg\min_{x\in C} D_\Psi(x, x_0). \tag{2}$$

i. (½ point) Show that the Bregman projection is well-defined.

ii. (½ point) Show that $\Pi_C(x) = x$ if and only if $x \in C$. Further, show that $\Pi_C \circ \Pi_C = \Pi_C$.

iii. (1 point) Show that, for all $x \in C$, $y \in \mathrm{dom}(\Psi)$, the following inequality holds:

$$\Big\langle \nabla\Psi(y) - \nabla\Psi(\Pi_C(y)), x - \Pi_C(y) \Big\rangle \le 0. \tag{3}$$

iv. (½ point) Deduce that, for all $x \in C$, $y \in \mathrm{dom}(\Psi)$, the following inequality holds:

$$D_\Psi(x, \Pi_C(y)) + D_\Psi(\Pi_C(y), y) \le D_\Psi(x, y). \tag{4}$$

v. (½ point) Show that, for $x, y, z \in \mathrm{dom}(\Psi)$, the following equality holds:

$$\langle \nabla\Psi(x) - \nabla\Psi(y), x - z\rangle = D_\Psi(x, y) + D_\Psi(z, x) - D_\Psi(z, y). \tag{5}$$

Solution:

i. The Bregman projection is well-defined as the function D_Ψ is strictly convex.

ii. $D_\Psi \ge 0$ and $D_\Psi(x, x_0) = 0$ if and only if $x = x_0$. The result follows.

iii. Note that we have $\nabla_x(D_\Psi(x, y)) = \nabla\Psi(x) - \nabla\Psi(y)$. Further, according to the convex optimality condition, we have $p = \Pi_C(y) = \arg\min_{z\in C} D_\Psi(z, y)$ if and only if for all $z \in C$, we have $\langle \nabla_x D_\Psi(p, y), z - p\rangle \ge 0$. Thus, we get the result.

iv. Denote $p = \Pi_C(y)$. We have

$$\begin{aligned} D_\Psi(x, p) + D_\Psi(p, y) - D_\Psi(x, y) = {} & \Psi(x) - \Big(\Psi(p) + \langle\nabla\Psi(p), x - p\rangle\Big) \\ & + \Psi(p) - \Big(\Psi(y) + \langle\nabla\Psi(y), p - y\rangle\Big) \\ & - \Psi(x) + \Big(\Psi(y) + \langle\nabla\Psi(y), x - y\rangle\Big) \\ = {} & -\langle\nabla\Psi(p), x - p\rangle + \langle\nabla\Psi(y), x - p\rangle \\ = {} & \langle\nabla\Psi(y) - \nabla\Psi(p), x - p\rangle \le 0 \end{aligned}$$

v. We have

$$\begin{aligned} D_\psi(x,y) &= \Psi(x) - \Big(\Psi(y) + \langle \nabla\Psi(y), x-y\rangle\Big) \\ D_\psi(z,x) &= \Psi(z) - \Big(\Psi(x) + \langle \nabla\Psi(x), z-x\rangle\Big) \\ D_\psi(z,y) &= \Psi(z) - \Big(\Psi(y) + \langle \nabla\Psi(y), z-y\rangle\Big) \\ D_\psi(x,y) + D_\psi(z,x) - D_\psi(z,y) &= -\langle \nabla\Psi(y), x-y\rangle - \langle \nabla\Psi(x), z-x\rangle + \langle \nabla\Psi(y), z-y\rangle \\ &= \langle \nabla\Psi(x) - \nabla\Psi(y), x-z\rangle \end{aligned}$$

(d) We now consider the mirror descent algorithm in the constrained case. Question marked with (♠) require more computation.

Assume that Ψ is ρ-convex, that C is a closed convex compact set, and f is L-Lipschitz with respect to $\|\cdot\|$. The mirror descent algorithm in the constrained case is defined as

$$y^{(k+1)} = \nabla\Psi^\star\Big(\nabla\Psi(x^{(k)}) - \gamma\nabla f(x^{(k)})\Big) \tag{6a}$$

$$x^{(k+1)} = \Pi_C(y^{(k+1)}). \tag{6b}$$

Let $x^\sharp \in \arg\min_{x\in C} f$, and $k \in \mathbb{N}$.

i. (½ point) Justify that $\nabla\Psi(y^{(k+1)}) = \nabla\Psi(x^{(k)}) - \gamma\nabla f(x^{(k)})$.

ii. (1 point) Show that the following inequality holds:

$$f(x^{(k)}) - f(x^\sharp) \le \frac{1}{\gamma}\Big\langle \nabla\Psi(x^{(k)}) - \nabla\Psi(y^{(k+1)}), x^{(k)} - x^\sharp\Big\rangle. \tag{7}$$

iii. (1 point) Using the results of c), deduce that the following inequality holds (♠):

$$f(x^{(k)}) - f(x^\sharp) \le \frac{1}{\gamma}\Big(D_\psi(x^\sharp, x^{(k)}) + D_\psi(x^{(k)}, y^{(k+1)}) - D_\psi(x^\sharp, x^{(k+1)}) - D_\psi(x^{(k+1)}, y^{(k+1)})\Big). \tag{8}$$

iv. (2 points) Assume that Ψ is ρ-convex, and show that the following inequality holds (♠):

$$D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)}) \le \Big\langle \nabla\Psi(y^{(k+1)}) - \nabla\Psi(x^{(k)}), x^{(k+1)} - x^{(k)}\Big\rangle - \frac{\rho}{2}\|x^{(k+1)} - x^{(k)}\|^2. \tag{9}$$

v. (1 point) Deduce that the following inequality holds:

$$D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)}) \le \gamma L\|x^{(k+1)} - x^{(k)}\| - \frac{\rho}{2}\|x^{(k+1)} - x^{(k)}\|^2. \tag{10}$$

vi. (½ point) Deduce that the following inequality holds:

$$D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)}) \le \frac{(\gamma L)^2}{2\rho}. \tag{11}$$

vii. (1 point) Deduce that the following inequality holds:

$$\sum_{k=1}^{K}(f(x^{(k)}) - f(x^\sharp)) \le \frac{D_\Psi(x^\sharp, x^{(1)})}{\gamma} + \gamma\frac{L^2 K}{2\rho}. \tag{12}$$

viii. (1 point) We denote $\bar{x}_K = \frac{1}{K}\sum_{k=1}^{K} x^{(k)}$. Show that

$$f(\bar{x}_K) - f(x^\sharp) \le \frac{D_\Psi(x^\sharp, x^{(1)})}{\gamma K} + \gamma\frac{L^2}{2\rho}. \tag{13}$$

Solution:

i. Direct consequence of a.v) and the definition of the mirror descent algorithm.

ii. By convexity of f, we have

$$\begin{aligned} & f(x^\sharp) \geq f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^\sharp - x^{(k)} \rangle \\ \Leftrightarrow\ & \langle \nabla f(x^{(k)}), x^{(k)} - x^\sharp \rangle \geq f(x^{(k)}) - f(x^\sharp) \end{aligned}$$

And as $\gamma \nabla f(x^{(k)}) = \nabla \Psi(x^{(k)}) - \nabla \Psi(y^{(k+1)})$, we get the result.

iii. By (5), we have

$$\langle \nabla \Psi(x^{(k)}) - \nabla \Psi(y^{(k+1)}), x^{(k)} - x^\sharp \rangle = D_\Psi(x^{(k)}, y^{(k+1)}) + D_\Psi(x^\sharp, x^{(k)}) - D_\Psi(x^\sharp, y^{(k+1)})$$

Further, as $x^{(k+1)} = \Pi_C(y^{(k+1)})$, we have, by (4), that

$$D_\Psi(x^\sharp, x^{(k+1)}) + D_\Psi(x^{(k+1)}, y^{(k+1)}) \leq D_\Psi(x^\sharp, y^{(k+1)}).$$

The result follows.

iv. We have

$$\begin{aligned} D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)}) =& \Psi(x^{(k)}) - \Big(\Psi(y^{(k+1)}) + \langle \nabla \Psi(y^{(k+1)}), x^{(k)} - y^{(k+1)} \rangle\Big) \\ & - \Psi(x^{(k+1)}) + \Big(\Psi(y^{(k+1)}) + \langle \nabla \Psi(y^{(k+1)}), x^{(k+1)} - y^{(k+1)} \rangle\Big) \\ =& \Psi(x^{(k)}) - \Psi(x^{(k+1)}) - \langle \nabla \Psi(y^{(k+1)}), x^{(k)} - x^{(k+1)} \rangle \end{aligned}$$

Further, as Ψ is ρ-convex, we have

$$\Psi(x^{(k+1)}) \geq \Psi(x^{(k)}) + \langle \nabla \Psi(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{\rho}{2} \|x^{(k+1)} - x^{(k)}\|^2$$

Thus

$$\begin{aligned} D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)}) \leq& \Psi(x^{(k)}) - \Big(\Psi(x^{(k)}) + \langle \nabla \Psi(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{\rho}{2} \|x^{(k+1)} - x^{(k)}\|^2\Big) \\ & - \langle \nabla \Psi(y^{(k+1)}), x^{(k)} - x^{(k+1)} \rangle \\ =& \langle \nabla \Psi(y^{(k+1)}) - \nabla \Psi(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle - \frac{\rho}{2} \|x^{(k+1)} - x^{(k)}\|^2 \end{aligned}$$

v. Recall that $\nabla \Psi(y^{(k+1)}) = \nabla \Psi(x^{(k)}) - \gamma \nabla f(x^{(k)})$, so

$$\langle \nabla \Psi(y^{(k+1)}) - \nabla \Psi(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle = \gamma \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$$

As f is L-Lipschitz, we have $\|\nabla f(x^{(k)})\|_* \leq L$, thus

$$\langle \nabla \Psi(y^{(k+1)}) - \nabla \Psi(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle \leq \gamma L \|x^{(k+1)} - x^{(k)}\|.$$

vi. The function $t \mapsto \gamma L t - \frac{\rho}{2} t^2$ is maximized for $t = \frac{\gamma L}{\rho}$, and its maximum value is $\frac{(\gamma L)^2}{2\rho}$.

vii. We sum inequality (8) for $k = 1$ to K and get

$$\sum_{k=1}^{K} (f(x^{(k)}) - f(x^\sharp)) \leq \frac{1}{\gamma} \Bigg(D_\Psi(x^\sharp, x^{(1)}) - \underbrace{D_\Psi(x^\sharp, x^{(K+1)})}_{\geq 0} + \sum_{k=1}^{K} \underbrace{D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)})}_{\leq \frac{(\gamma L)^2}{2\rho} \text{ by (11)}} \Bigg)$$

viii. By convexity of f, we have

$$f(\bar{x}_K) - f(x^\sharp) \leq \frac{1}{K} \sum_{k=1}^{K} f(x^{(k)}) - f(x^\sharp)$$

The result follows from the previous inequality.