MATH 6222 LECTURE NOTE 2: UNCONSTRAINED CONVEX OPTIMIZATION AND GRADIENT DESCENT METHODS

JINGRONG WEI

ABSTRACT. In this note, we discuss the existence and uniqueness of the solution to unconstrained convex optimization problems. We present convergence analysis for gradient flow and gradient descent methods. The main reference is [1].

CONTENTS

Consider the unconstrained convex minimization problem:

$$\min_{x\in\mathbb{R}^n} f(x),$$

where f is a smooth convex function ($f \in \mathcal{S}^{1,1}_{\mu,L}$).

1. EXISTENCE AND UNIQUENESS OF SOLUTION

A point x^* is a *global minimizer* if $f(x^*) \le f(x)$ for all $x \in \mathbb{R}^n$. A point x^* is a local minimizer if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x^*) \le f(x)$ for all $x \in \mathcal{N}$. The minimizer is a strict minimizer if the inequality is strict.

A function f is called *lower semicontinuous* at $x \in V$ if

$$f(x) \le \liminf_{k\to\infty} f(x_k)$$

for every sequence $\{x_k\} \subset V$ with $x_k \to x$. f is lower semicontinuous if it is lower semicontinuous at each $x \in V$.

A function f is called *coercive* if for all sequence $\{x_k\}$ with $\|x_k\| \to \infty$, we have $\lim_{k\to\infty} f(x_k) = \infty$.

Proposition 1.1. *Suppose V is non-empty and closed and $f : V \to \mathbb{R}$ is lower-semicontinuous and coercive. Then f has a global minimum over V.*

Date: January 22, 2025.

Proof. We may assume that $f(x) < \infty$ for some $x \in V$. Then $f^* = \inf_{x \in V} f(x) < \infty$. Let $\{x_k\} \subset V$ be a sequence such that $\lim_{k \to \infty} f(x_k) = f^* < \infty$. Then since f is coercive, $\{x_k\}$ is bounded. Then there exists a subsequence x_{k_j} converging to a point x^*. Since V is closed, $x^* \in V$. Then

$$f^* = \lim_{k \to \infty} f(x_k) = \lim_{j \to \infty} f\left(x_{k_j}\right) \geq f(x^*)$$

Therefore, x^* is a global minimmum of f over V. □

Recall that we have first-order necessary conditions, second-order necessary/sufficient conditions to characterize local minimizers.

Theorem 1.2 (First-Order Necessary Conditions)**.** *If x^* is a local minimizer and f is continuously differentiable in an open neighborhood of x^*, then $\nabla f(x^*) = 0$.*

Theorem 1.3 (Second-Order Necessary Conditions)**.** *If x^* is a local minimizer of f and $\nabla^2 f$ exists and is continuous in an open neighborhood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.*

Theorem 1.4 (Second-Order Sufficient Conditions)**.** *Suppose that $\nabla^2 f$ is continuous in an open neighborhood of x^* and that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of f.*

The most important properties of convex functions is that we could find global minimizers under this setting,

Theorem 1.5. *If $f \in \mathcal{S}^1(\mathbb{R}^n)$ and $\nabla f(x^*) = 0$ then x^* is the global minimizer of $f(\cdot)$ on $\mathbb{R}^n$.*

Proof. In view of definition, for any $x \in \mathbb{R}^n$ we have

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle = f(x^*)$$

□

Remark 1.6. *The minimizer may not be unique. However, the solution forms a convex set.*

Theorem 1.7. *If $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$ and $\nabla f(x^*) = 0$, then*

$$f(x) \geq f(x^*) + \frac{1}{2}\mu \|x - x^*\|^2$$

for all $x \in \mathbb{R}^n$.

Remark 1.8. *If the solution of minimizing a strongly convex function exists, then it is the unique solution.*

2. Gradient Descent Methods

In this section, we present convergence analysis for gradient descent methods. We start with introducing gradient flow.

2.1. **Gradient Flow.** The simplest dynamical system associated with the unconstrained minimization problem is the gradient flow:

$$x'(t) = -\nabla f(x(t)), \tag{1}$$

with the initial condition $x(0) = x_0$. Assume $f \in \mathcal{S}^1_\mu$ with $\mu > 0$ and let x^* be the global minimizer of f. Our primary interest lies in analyzing the convergence of $x(t)$ to x^* as $t \to \infty$.

In order to study the stability of an equilibrium x^* of a autonomous dynamical system defined by a vector field $\mathcal{G} : V \to V$:

$$x' = \mathcal{G}(x(t)), \tag{2}$$

Lyapunov introduced the so-called Lyapunov function $\mathcal{E}(x)$ [3, 4], which is nonnegative and the equilibrium point x^* satisfies $\mathcal{E}(x^*) = 0$ and the Lyapunov condition: $-\nabla\mathcal{E}(x) \cdot \mathcal{G}(x)$ is locally positive near the equilibrium point x^*. That is the flow $\mathcal{G}(x)$ may not be the perfect $-\nabla\mathcal{E}(x)$ direction but contains positive component in that direction. Then the (local) decay property of $\mathcal{E}(x)$ along the trajectory $x(t)$ of the autonomous system (2) can be derived immediately

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(x(t)) = \nabla\mathcal{E}(x) \cdot x'(t) = \nabla\mathcal{E}(x) \cdot \mathcal{G}(x) < 0.$$

To further establish the convergence rate of $\mathcal{E}(x(t))$, Chen and Luo [1] introduced the strong Lyapunov condition: suppose there exists a compact subset $W \subseteq V$, a positive constant $c > 0$, a constant $q \geqslant 1$, and a function $p(x) : V \to \mathbb{R}$, such that $\mathcal{E}(x) \geqslant 0$ and

$$-\nabla\mathcal{E}(x) \cdot \mathcal{G}(x) \geqslant c\,\mathcal{E}^q(x) + p^2(x) \quad \forall\, x \in W. \tag{3}$$

Under these conditions, we call $\mathcal{E}$ a locally strong Lyapunov function (if $W \subset V$) or a globally strong Lyapunov function (if $W = V$).

Theorem 2.1. *Assume that $\mathcal{E}(x)$ satisfies the strong Lyapunov property (3). If the trajectory $x(t)$ of (2) satisfies $\{x(t) : t \geqslant 0\} \subset W$, then for all $t \geqslant 0$,*

$$\mathcal{E}(x(t)) + \int_0^t e^{c(s-t)}\|p(x(s))\|^2\,\mathrm{d}s \leqslant \mathcal{E}_0 \exp(-c\,t), \quad \textit{for } q = 1. \tag{4}$$

and

$$\mathcal{E}(x(t)) \leqslant \Big((q-1)c\,t + \mathcal{E}_0^{1-q}\Big)^{1/(1-q)}, \quad \textit{for } q > 1, \tag{5}$$

where $\mathcal{E}_0 = \mathcal{E}(x(0))$.

Proof. By assumption, for all $t \geqslant 0$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(x(t)) &= \nabla\mathcal{E}(x(t)) \cdot x'(t) = \nabla\mathcal{E}(x(t)) \cdot \mathcal{G}(x(t)) \\ &\leqslant -c\,\mathcal{E}^q(x(t)) - \|p(x(t))\|^2\end{aligned} \tag{6}$$

For the case $q = 1$, integrating (6) immediately gives the desired result.

Now consider $q > 1$. From (6), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^{1-q}(x(t)) = (1-q)\frac{\mathcal{E}'(x(t))}{\mathcal{E}^q(x(t))} \geqslant c(q-1).$$

Integrating this inequality yields

$$\mathcal{E}^{1-q}(x(t)) - \mathcal{E}^{1-q}(x(0)) \geqslant c(q-1)t, \quad t \geqslant 0.$$

Rearranging gives

$$\mathcal{E}(x(t)) \leqslant \left((q-1)c\,t + \mathcal{E}_0^{1-q}\right)^{1/(1-q)},$$

which completes the proof. □

Furthermore if $\|x - x^*\|^2 \le C\mathcal{E}(x)$, then we can derive the stability of x^* from the decay of Lyapunov function $\mathcal{E}(x)$.

A natural choice for the Lyapunov function is the optimality gap:

$$\mathcal{E}(x) = f(x) - f(x^\star). \tag{7}$$

Direct computation gives

$$-\nabla\mathcal{E}(x)\cdot\mathcal{G}(x) = \|\nabla f(x)\|_*^2 \ge \mu\,\mathcal{E}(x) + \frac{1}{2}\|\nabla f(x)\|_*^2.$$

For the inequality, we use the bound for $f \in \mathcal{S}_\mu^1$

$$\|\nabla f(x)\|_*^2 = \|\nabla f(x) - \nabla f(x^*)\|_*^2 \ge 2\mu D_f(x, x^*) = 2\mu\mathcal{E}(x).$$

Notice we split the $\|\nabla f(x)\|_*^2$ to have an additional positive term $p^2 = \|\nabla f(x)\|_*^2/2$.

We can consider alternative candidates for strong Lyapunov functions besides the optimality gap in (7). Two notable examples are presented below. One is the squared distance to the minimizer:

$$\mathcal{E}(x) = \frac{1}{2}\|x - x^\star\|^2.$$

Using inequality [5, Theorem 2.1.12]

$$M_{\nabla f}(x, x^*) \ge \frac{\mu L}{L+\mu}\|x - x^*\|^2 + \frac{1}{L+\mu}\|\nabla f(x)\|_*^2,$$

we obtain:

$$-\nabla\mathcal{E}(x)\cdot\mathcal{G}(x) = \langle x - x^\star, \nabla f(x)\rangle \ge \frac{2\mu L}{L+\mu}\mathcal{E}(x) + \frac{1}{L+\mu}\|\nabla f(x)\|_*^2.$$

Another effective candidate is a combination of the optimality gap and squared distance:

$$\mathcal{E}(x) = f(x) - f(x^\star) + \frac{\mu}{2}\|x - x^\star\|^2.$$

Direct computation gives

$$-\nabla\mathcal{E}(x)\cdot\mathcal{G}(x) = \|\nabla f(x)\|_*^2 + \mu\langle x - x^\star, \nabla f(x)\rangle \ge \mu\mathcal{E}(x) + \|\nabla f(x)\|_*^2.$$

We summarize the results in the following proposition.

Proposition 2.2. *Assume $f \in \mathcal{S}_{\mu,L}^{1,1}$ with $0 < \mu \le L \le \infty$. For the gradient flow $x'(t) = -\nabla f(x(t))$, we have the following strong Lyapunov functions:*

$$\mathcal{E}(x) = f(x) - f(x^\star), \quad -\nabla\mathcal{E}(x)\cdot\mathcal{G}(x) \ge \mu\mathcal{E}(x) + \frac{1}{2}\|\nabla f(x)\|_*^2, \tag{8}$$

$$\mathcal{E}(x) = \frac{1}{2}\|x - x^\star\|^2, \quad -\nabla\mathcal{E}(x)\cdot\mathcal{G}(x) \ge \frac{2\mu L}{L+\mu}\mathcal{E}(x) + \frac{1}{L+\mu}\|\nabla f(x)\|_*^2, \tag{9}$$

$$\mathcal{E}(x) = f(x) - f(x^\star) + \frac{\mu}{2}\|x - x^\star\|^2, \quad -\nabla\mathcal{E}(x)\cdot\mathcal{G}(x) \ge \mu\mathcal{E}(x) + \|\nabla f(x)\|_*^2. \tag{10}$$

Consequently Theorem 2.1 guarantees the exponential decay $\mathcal{O}(e^{-ct})$ for both $\mathcal{E}(x)$ and $\|\nabla f(x)\|_*$ along the trajectory of the gradient flow (1).

When verifying the strong Lyapunov property, we retain an extra positive term $p^2 = C\|\nabla f(x)\|_*^2$. This additional term is useful for analyzing the gradient descent method.

When $\mu = 0$, the strong Lyapunov properties mentioned earlier degenerate. Define the sub-level set of f for a given constant value c as:

$$S_c(f) = \{x\colon f(x) \leq c\}.$$

Since f is convex, the sub-level set $S_c(f)$ is also convex. The set of minimizers of f, where it attains its minimum value $f_{\min}$, can be expressed as $S_{f_{\min}}(f)$.

Lemma 2.3. *Let f be convex and coercive. For a given finite value f_0, there exists a constant R_0 such that*

$$\max_{x^\star \in \operatorname{argmin} f} \max_{x \in S_{f_0}} \|x - x^\star\| \leq R_0. \tag{11}$$

Proof. If S_{f_0} were unbounded, we could construct a sequence $\{x_n\}$ such that $f(x_n) \leq f_0$ but $\|x_n\| > n$ for $n = 1, 2, \ldots$, which would contradict the coercivity of f.

Additionally, $\operatorname{argmin} f \subseteq S_{f_0}$, so $\operatorname{argmin} f$ is also bounded. Thus, (11) holds. □

Proposition 2.4. *Let f be convex and coercive. For $\mathcal{G}(x) = -\nabla f(x)$, we have the following strong Lyapunov function $\mathcal{E}(x) = f(x) - f(x^\star)$, where $x^\star$ is an arbitrary but fixed point in the minimum set* $\operatorname{argmin} f$*,*

$$-\nabla \mathcal{E}(x) \cdot \mathcal{G}(x) \geq \frac{1}{R_0^2}\mathcal{E}^2(x) \quad \forall x \in S_{f_0}(f), \tag{12}$$

where R_0 is defined by (11) *and $f_0 = f(x_0)$. Consequently the trajectory of the gradient flow $x(t)$ satisfies*

$$f(x(t)) - f(x^\star) \leq \frac{1}{R_0^2 t + C}, \quad \forall t > 0. \tag{13}$$

Proof. For $\mathcal{E}(x) = f(x) - f(x^\star)$ assuming coercivity and convexity, we have

$$f(x) - f(x^\star) \leq \langle \nabla f(x), x - x^\star \rangle \leq R_0 \|\nabla f(x)\|_* \quad \forall x \in S_{f_0}(f). \tag{14}$$

Thus, the strong Lyapunov property (12) follows by

$$-\nabla \mathcal{E}(x) \cdot \mathcal{G}(x) = \|\nabla f(x)\|_*^2 \geq \frac{1}{R_0^2}\mathcal{E}^2(x) \quad \forall x \in S_{f_0}(f).$$

Since $-\nabla \mathcal{E}(x) \cdot \mathcal{G}(x) = \|\nabla f(x)\|_*^2 \geq 0$, the trajectory of the gradient flow $x(t)$ satisfies $x(t) \in S_{f_0}(f)$. By applying Theorem 2.1, we conclude that the optimality gap $f(x(t)) - f(x^\star)$ decays at the sublinear rate $O(1/t)$ along the trajectory of the gradient flow. □

2.2. **Gradient Descent Methods.** With the strong Lyapunov property, the convergence of the implicit Euler method is straightforward.

Theorem 2.5. *Assume $\mathcal{E}(x)$ is a convex Lyapunov function satisfying the strong Lyapunov property for some $c > 0$:*

$$-\nabla \mathcal{E}(x) \cdot \mathcal{G}(x) \geqslant c\, \mathcal{E}(x), \quad \forall x \in V.$$

Let $\{x_k\}$ be the sequence generated by the implicit Euler method, starting from a given x_0, for $k = 0, 1, \ldots$

$$x_{k+1} - x_k = \alpha_k \mathcal{G}(x_{k+1}).$$

Then, for $k \geq 0$, the sequence satisfies the linear contraction:

$$\mathcal{E}(x_{k+1}) \leqslant \frac{1}{1 + c\,\alpha_k}\mathcal{E}(x_k).$$

Proof. For brevity, denote $\mathcal{E}_k = \mathcal{E}(x_k)$. Then

$$\begin{aligned} \mathcal{E}_{k+1} - \mathcal{E}_k &\leqslant (\nabla \mathcal{E}(x_{k+1}), x_{k+1} - x_k) && \text{convexity of } \mathcal{E} \\ &= \alpha_k (\nabla \mathcal{E}(x_{k+1}), \mathcal{G}(x_{k+1})) && \text{implicit Euler method} \\ &\leq -c\, \alpha_k \mathcal{E}_{k+1}. && \text{strong Lyapunov property at } x_{k+1} \end{aligned}$$

□

When $\mathcal{G}(x) = -\nabla f(x)$, the implicit Euler method applied to the gradient flow:

$$x_{k+1} = x_k - \alpha \nabla f(x_{k+1}) \tag{15}$$

can be reformulated as:

$$x_{k+1} = \operatorname{prox}_{\alpha f}(x_k) := \arg\min_x \left\{ f(x) + \frac{1}{2\alpha}\|x - x_k\|^2 \right\}, \tag{16}$$

which is known as the proximal point algorithm (PPA) [2, 6].

Next, we present the convergence analysis for the explicit Euler method applied to the gradient flow, which corresponds to the gradient descent method.

Theorem 2.6. *Assume $f \in \mathcal{S}^{1,1}_{\mu,L}$ with $0 < \mu \leq L < \infty$. Let $\{x_k\}$ be the sequence generated by*

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k). \tag{17}$$

For $\alpha_k \leq 2/(L+\mu)$, we have

$$\mathcal{E}_{k+1} \leq (1 - \mu\alpha_k)\mathcal{E}_k,$$

where $\mathcal{E}(x) = f(x) - f(x^\star) + \frac{\mu}{2}\|x - x^\star\|^2$ and $\mathcal{E}_k = \mathcal{E}(x_k)$.

The optimal value $\alpha_k = 2/(L+\mu)$ gives

$$\mathcal{E}_{k+1} \leq \frac{L-\mu}{L+\mu}\,\mathcal{E}_k,$$

and the quasi-optimal value $\alpha_k = 1/L$ gives

$$\mathcal{E}_{k+1} \leq (1 - \mu/L)\mathcal{E}_k.$$

Proof. As $f \in \mathcal{S}^{1,1}_{\mu,L} \subset \mathcal{S}^1_\mu$, we have verified the strong Lyapunov property in (10) with an additional term $\|\nabla f(x)\|^2$. Note that $\mathcal{E} \in \mathcal{S}^{1,1}_{2\mu,L+\mu}$. Using the definition of Bregman divergence, the upper bound of $D_{\mathcal{E}}$, and the strong Lyapunov condition at x_k, we have

$$\begin{aligned} \mathcal{E}_{k+1} - \mathcal{E}_k &= \langle \nabla \mathcal{E}(x_k), x_{k+1} - x_k \rangle + D_{\mathcal{E}}(x_{k+1}, x_k) \\ &\leq -\alpha_k \langle \nabla \mathcal{E}(x_k), \nabla f(x_k) \rangle + \frac{L+\mu}{2}\|x_{k+1} - x_k\|^2 \\ &\leq -\mu\, \alpha_k \mathcal{E}_k - \alpha_k \left(1 - \frac{L+\mu}{2}\alpha_k\right) \|\nabla f(x_k)\|_*^2. \end{aligned}$$

For $\alpha_k \leq 2/(L+\mu)$, we have $\mathcal{E}_{k+1} - \mathcal{E}_k \leq -\mu\, \alpha_k \mathcal{E}_k$, and the linear convergence follows. □

One can also choose

$$\mathcal{E}(x) = f(x) - f(x^\star) \quad \text{or} \quad \mathcal{E}(x) = \frac{1}{2}\|x - x^\star\|^2,$$

and prove the linear convergence of the gradient descent method. Here, we present a sufficient decay property of the function values.

Proposition 2.7. *Assume $f \in \mathcal{S}_{\mu,L}^{1,1}$ with $0 < \mu \leq L < \infty$. For the gradient descent method with $\alpha = 1/L$:*

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k),$$

we have the function decay estimate

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\left(\|\nabla f(x_{k+1})\|_*^2 + \|\nabla f(x_k)\|_*^2\right). \tag{18}$$

Consequently,

$$f(x_{k+1}) - f(x^\star) \leq \frac{L-\mu}{L+\mu}\big(f(x_k) - f(x^\star)\big). \tag{19}$$

Proof. Using the Bregman divergence and the identity of squares, we expand the difference:

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= \langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - D_f(x_k, x_{k+1}) \\ &= -\alpha\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle - D_f(x_k, x_{k+1}) \\ &= -\frac{\alpha}{2}\left(\|\nabla f(x_{k+1})\|_*^2 + \|\nabla f(x_k)\|_*^2\right) \\ &\quad + \frac{\alpha}{2}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|_*^2 - D_f(x_k, x_{k+1}). \end{aligned}$$

Then, using the bound $D_f(x_k, x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2$, we cancel the positive term, obtaining (18).

Finally, applying the inequality $f(x) - f(x^\star) \leqslant \frac{1}{2\mu}\|\nabla f(x)\|_*^2$ to relate the terms on the right-hand side to the optimality gap, we derive (19). □

Now consider the case $\mu = 0$ and introduce a perturbed objective

$$f_\epsilon(x) = f(x) + \frac{\epsilon}{2}\|x\|^2,$$

with some $\epsilon > 0$. Then it is clear that

$$f_\epsilon(x_\epsilon^*) \leq f_\epsilon(x^*) \quad\Longrightarrow\quad -f(x^*) \leq -f_\epsilon(x_\epsilon^*) + \frac{\epsilon}{2}\|x^*\|^2, \tag{20}$$

where x_ϵ^* denotes the unique global minimizer of f_ϵ and $x^* \in \operatorname{argmin} f$ is bounded, i.e., $\|x^*\| < \infty$. Since $f_\epsilon \in \mathcal{S}_{\epsilon,L+\epsilon}^{1,1}$, then applying gradient descent with $\alpha = 1/(L+\epsilon)$, and by Theorem 2.6 we have that

$$f_\epsilon(x_k) - f_\epsilon(x_\epsilon^*) + \frac{\epsilon}{2}\|y_k - x_\epsilon^*\|^2 \leq (1 - \epsilon/(L+\epsilon))^{-k}\,\mathcal{E}_\epsilon(x_0, y_0),$$

where $\mathcal{E}_\epsilon(x,y) = f_\epsilon(x) - f_\epsilon(x_\epsilon^*) + \frac{\epsilon}{2}\|y - x_\epsilon^*\|^2$.

Combining this with (20) yields

$$\begin{aligned} f(x_k) - f(x^*) &\leq f_\epsilon(x_k) - f_\epsilon(x_\epsilon^*) + \frac{\epsilon}{2}\left(\|x^*\|^2 - \|x_k\|^2\right) \\ &\leq \frac{\epsilon}{2}\|x^*\|^2 + (1 - \epsilon/(L+\epsilon))^{-k}\,\mathcal{E}_\epsilon(x_0, y_0). \end{aligned}$$

Note that both $\|x^*\|$ and $\mathcal{E}(x_0, y_0)$ are bounded constants. To achieve the accuracy $f(x_k) - f(x^*) \leq O(\epsilon)$, the number of iterations is bounded by

$$(1 - \epsilon/L)^{-k} = \epsilon \quad\Longrightarrow\quad k = \frac{|\ln\epsilon|}{\ln(1+\epsilon/L)} \sim \frac{L}{\epsilon}|\ln\epsilon|.$$

Compared with the complexity $\mathcal{O}(\epsilon^{-1})$, the log factor $|\ln \epsilon|$ is negligible. This gives complexity of gradient descent methods for convex problems. However, for practical performance, using the fixed small constant ϵ may not be efficient than dynamically changing sequence $\epsilon_k \to 0$.

Accelerated gradient methods achieve a linear convergence rate of $\mathcal{O}(1 - \sqrt{\mu/L})^k$ for strongly convex problems and a sublinear rate of $\mathcal{O}(1/k^2)$ for convex problems. These are optimal rates for first-order (using gradients) iterative methods. For some specific optimal methods, see [5, 1].

References

[1] L. Chen and H. Luo. A unified convergence analysis of first order convex optimization methods via strong Lyapunov functions. *arXiv preprint arXiv:2108.00132*, 2021. 1, 3, 8

[2] O. Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM journal on control and optimization*, 29(2):403–419, 1991. 6

[3] W. M. Haddad and V. Chellaboina. *Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach*. Princeton University Press, 2008. 3

[4] H. K. Khalil. *Nonlinear systems; 3rd ed.* Prentice-Hall, Upper Saddle River, NJ, 2002. 3

[5] Y. Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018. 4, 8

[6] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976. 6