A HAMILTONIAN PERSPECTIVE ON SHERMAN'S AREA-CONVEXITY ALGORITHM

JUNFEI SUN

Abstract. This paper aims at two major things: (1) trying to provide a Hamiltonian and symplectic geometry perspective onto Sherman's area-complexity algorithm and serving as a starter for a more general symplectic duality theory. This perspective is based on the construction of an analogy of Fenchel's conjugate in terms of symplectic forms and taking it as the Hamiltonian. (2) Precisely closing the gap between more general optimization methods and Sherman's area convexity algorithm and their corresponding results for convergence rate by introducing a variant of dual extrapolation and linking it with Sherman's algorithm. We first introduce basic theories regarding Hamiltonian systems and Liouville's theorem in Section 2. We then tackle (1) in Section 3 and finally (2) in Section 4.

Contents

1. Introduction

Sherman's paper has introduced an algorithm that's based on using an area-convex regularizer instead of a strongly-convex regularizer. Such softening of the requirement on the regularizer enables the breaking of l_∞-barrier and achieves an accelerated running time for multi-commodity flow problems. However, this algorithm is presented in a very ad-hoc way. In this paper, we try to unwrap this algorithm in two ways. Our first goal is to try to interpret this algorithm in terms of a Hamiltonian perspective. Along the same line, we believe that this algorithm

potentially points to a duality theory using symplectic geometry for saddle point problems. Secondly, we want to tackle the ad-hoc nature of this algorithm by introducing a variant of dual-extrapolation and showing how this algorithm is a specific case for this variant of dual-extrapolation.

To the stated ends, we will first state theories established for Hamiltonian systems in Section 2. We begin by giving the definition of Hamiltonian systems and Hamiltonian vector fields, followed by stating the important property of energy conservation of the flow of Hamiltonian vector fields in Section 2.1. We will later show that this property is crucial for the guarantee of Sherman's area convexity algorithm and a variant of dual extrapolation in Section 3 and Section 4. We will then introduce how the stationary principle induces Hamiltonian systems with Euler-Langrange equations as a mediator in Section 2.2. We close Section 2 by introducing canonical transformations and stating an important equivalent characterization of canonical transformations that's related to Hamiltonian systems. Finally, we will prove Liouvilles' theorem using Stoke's theorem and raise that this proof can be extended to all vector fields with divergence 0 on a contractible manifold. Noticeably, we won't explicitly link Liouvilles' theorem with Sherman's algorithm and dual extrapolation in later sections but we conjecture that Liouvilles' theorem should appear when an extensive duality theory based on symplectic conjugate is developed (which we get started in Section 3).

In Section 3, we first reduce the general bi-linear saddle point problem into a self-dual saddle point problem. We then introduce Sherman's algorithm based on area convexity (Algorithm 1) that finds an approximate solution for a self-dual saddle point problem in accelerated running time. After that, we construct the symplectic conjugate in analogy to Fenchel's conjugate and prove a version of conjugate correspondence between smoothness and area convexity. More generally, this construction can be the key to a symplectic duality theory. Finally, we close Section 3 by posing the crucial observation that Algorithm 1 can be seen as a variation of the discretization of the Hamiltonian system with the symplectic conjugate as the Hamiltonian. Additionally, the property of energy conservation of moving along the flow of the Hamiltonian field suffices to guarantee the convergence rate of Algorithm 1. However, we will show that this variation of discretization doesn't fall into a nice category of numerical methods that have nice long-term energy behavior: symplectic numerical methods.

Finally, in Section 4, we tackle the ad-hoc nature of Algorithm 1 by linking it with dual-extrapolation as a more general optimization method. We will first see that Algorithm 1 is exactly a variation of the dual extrapolation. We then turn to the convergence analysis of this variation of dual extrapolation. This analysis involves using the concept of relaxed-relative lipschitzness as a generalization of both area-convexity and relaxed lipschitzness. This relaxation of the condition of the function introduces a constant factor to the rate of convergence of this variation of dual extrapolation when compared to that of ordinary dual extrapolation. Finally, we introduce the same Hamiltonian perspective to a specific case of dual extrapolation and note that the guarantee of convergence of dual extrapolation can also be seen as the result of the conservation of energy.

2. Hamiltonian

2.1. **Hamiltonian systems and conservation of energy.** In this section, we introduce a very important system at the root of explaining a lot of mechanical systems and an important property of such a system. If we are given a C^1 map $f : TM \to \mathbb{R}$ where M is an $n-$dimensional submanifold of $\mathbb{R}^n$, we can construct the vector field H_f on TM as the following:

$$H_f = \sum_{j=1}^{n} [\frac{\partial f}{\partial p_j}\frac{\partial}{\partial q_j} - \frac{\partial f}{\partial q_j}\frac{\partial}{\partial p_j}] \tag{2.1}$$

where (q, p) refers to the coordinates of TM. H_f is called the Hamiltonian vector field and f is its Hamiltonian.

Notice that we are writing H_f in the differential operator form, i.e. given a function g

$$H_f g(x) = \lim_{h\to 0} \frac{g(F^h_{H_f}(x)) - g(x)}{h} \tag{2.2}$$

where $F^h_{H_f}(x)$ denotes the flow of H_f at time h with starting point at x.

Therefore, for a vector field $V(x) = \sum a_i(x)\frac{\partial}{\partial x_i}$, we have that for any $1 \le i \le n$, $Vx_i = a_i(x)$. This implies that

$$a_i(x) = \lim_{h\to 0} \frac{(F^h_V(x))_i - x_i}{h} = D_i F^0_V(x) = D_i x = V(x)_i. \tag{2.3}$$

Applying this conclusion to the above Hamiltonian vector field, we obtain that

$$H_f(x) = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} Df(x). \tag{2.4}$$

Therefore, when we consider the integral curve of H_f, i.e., the solution to the ODE:

$$\dot{x} = H_f(x), \quad x(0) = x_0, \tag{2.5}$$

it can be seen that the integral curve is characterized by the following set of equations

$$\begin{cases} \dot{q} &= D_p f(q, p) \\ \dot{p} &= -D_q f(q, p) \end{cases} \tag{2.6}$$

where we let $x = (q, p)$ and $(q(0), p(0)) = (q_0, p_0) := x_0$. (2.6) are called the Hamiltonian equations. More generally, for Hamiltonian equations, f can also explicitly depend on time, under which scenario we say that the Hamiltonian equations are time-dependent. We remark that in physics settings, q is usually considered as the general position and p the general momentum, and the integral curve of the Hamiltonian is viewed as the trajectory of the mechanical behavior of the target object in the phase space of (q, p). We will later see in section 2.2 how Hamiltonian systems naturally arise out of the stationary variation principle which is considered axiomatic in Hamiltonian mechanics.

One important property of the integral curve of the Hamiltonian vector field, or the solution to time-independent Hamiltonian equations, is that it lies on a level set of f. If we consider f as some notion of 'energy', the integral curve of the Hamiltonian preserves energy.

Proposition 2.7. *Given a Hamiltonian vector field H_f, the integral curve $x(t) = (q(t), p(t))$ of H_f preserves f, i.e., for any t_1, t_2, we have that $f(x(t_1)) = f(x(t_2))$*

Proof. Let $E(t) = f(q(t), p(t))$. We have that

$$\begin{aligned}\dot{E}(t) &= Df(q(t), p(t))Dx(t) = D_q f(q,p)\dot{q}(t) + D_p f(q,p)\dot{p}(t) \\ &= -\dot{p}(t)\dot{q}(t) + \dot{q}(t)\dot{p}(t) = 0\end{aligned}$$

which implies that E is preserved. □

Remark 2.8. We will later argue in section 3 and section 4 that this conservation of energy is actually crucial for the guarantee of the rate of convergence of the algorithm proposed in Sherman's paper and also a variation of dual extrapolation under a specific case.

2.2. **Stationary action principle.** With the concept of Hamiltonian systems established, we'll now turn our attention to how are these systems generated in Hamiltonian mechanics. The key is the relationship between the stationary action principle and Hamiltonian systems. To make sense of this relationship, we first give the definition of action:

Definition 2.9. Let M be an $n-$dimensional submanifold of $\mathbb{R}^n$. Let $x, y \in M$ and $\mathcal{P}$ be the family of smooth funtions $u : [a, b] \to M$ such that $u(a) = x$ and $u(b) = y$. Given a smooth function $L : TM \times \mathbb{R} \to \mathbb{R}$, we let $I : \mathcal{P} \to \mathbb{R}$ be defined as:

$$I(u) = \int_a^b L(u(t), \dot{u}(t), t)dt \tag{2.10}$$

and we call I the action w.r.t L.

We say that the action is stationary if for any smooth family $\{u_s\}$ in $\mathcal{P}$ with $u_0 = u$, we have that

$$\frac{d}{ds} I(u_s)|_{s=0} = 0. \tag{2.11}$$

Remark 2.12. An action being stationary suggests $I(u + hn(t)) - I(u) = o(h)$ where $n(t)$ is some small smooth function that vanishes on $t = a, b$ that depends on u_s. Intuitively, this means making any small variation to u won't change the action.

An important result for linking Stationary action and Hamiltonians is that an action is stationary if and only if it follows the so-called Euler-Lagrange equation:

Theorem 2.13. *The action I w.r.t L is stationary if and only if*

$$D_u L - \frac{d}{dt}(D_{\dot{u}} L) = 0. \tag{2.14}$$

Proof. Firstly, notice that

$$\begin{aligned}\frac{d}{ds} I(u_s)|_{s=0} &= \frac{d}{ds}\int_a^b L(u_s(t), \dot{u_s}(t), t)dt|_{s=0} \\ &= \int_a^b \frac{d}{ds} L(u_s(t), \dot{u_s}(t), t)|_{s=0} dt \text{ (by smoothness of } L \text{ and } u_s) \\ &= \int_a^b D_u L(u(t), \dot{u}(t), t)(\frac{d}{ds} u_s(t)|_{s=0}) + D_{\dot{u}} L(u(t), \dot{u}(t), t)(\frac{d}{ds}\dot{u_s}(t)|_{s=0})dt.\end{aligned}$$

Integrating the last term by parts, we obtain:

$$\begin{aligned}
&\int_a^b D_{\dot{u}}L\,(\frac{d}{ds}\dot{u}_s(t)|_{s=0})dt \\
=&D_{\dot{u}}L\,(\frac{d}{ds}(u_s(b)-u_s(a))|_{s=0}) - \int_a^b (\frac{d}{dt}D_{\dot{u}}L)\,(\frac{d}{ds}u_s(t)|_{s=0})dt \\
=&-\int_a^b (\frac{d}{dt}(D_{\dot{u}}L))\,(\frac{d}{ds}u_s(t)|_{s=0})dt \text{ (since } u_s(a)\equiv x, u_s(b)\equiv y \text{ by definition).}
\end{aligned}$$

Therefore, we have

$$\frac{d}{ds}I(u_s)|_{s=0} = \int_a^b (D_uL - \frac{d}{dt}(D_{\dot{u}}L))\,(\frac{d}{ds}u_s(t)|_{s=0})dt.$$

Now, I is stationary if and only if $\frac{d}{ds}I(u_s)|_{s=0}=0$ for any smooth family of u_s, which is equivalent to $\int_a^b(D_uL-\frac{d}{dt}(D_{\dot{u}}L))\,(\frac{d}{ds}u_s(t)|_{s=0})dt=0$ for any smooth family of u_s. And that is equivalent to $D_uL-\frac{d}{dt}(D_{\dot{u}}L)=0$, which concludes the proof.

□

With theorem 2.13 established, we now try to link the stationary action principle and Hamiltonians by linking the Euler-Lagrange equation and Hamiltonian equations.

Proposition 2.15. *Let $H(q,p,t)$ be obtained from Legrenge transformation from $L(q,\dot{q},t)$, i.e., $H(q,p,t)=\sum_{i=1}^n p_i\dot{q}_i - L(q,\dot{q},t)$ where $p:=D_{\dot{q}}L$, then*

$$D_qL - \frac{d}{dt}(D_{\dot{q}}L)=0 \iff \begin{cases} \dot{q} &= D_pH(q,p,t) \\ \dot{p} &= -D_qH(q,p,t) \end{cases}.$$

Proof. Notice that $\dot{q}=D_pH(q,p,t)$ follows directly from Legrenge transformation. Therefore, we only need to show that $D_qL-\frac{d}{dt}(D_{\dot{q}}L)=0 \iff \dot{p}=-D_qH(q,p,t)$.

To see this, we will first show that $D_qH=-D_qL$. Notice

$$\begin{aligned}
dH &= \sum_{i=1}^n p_i d\dot{q}_i + \dot{q}_i dp_i - dL \\
&= \sum_{i=1}^n p_i d\dot{q}_i + \dot{q}_i dp_i - \sum_{i=1}^n D_{\dot{q}_i}L d\dot{q}_i - \sum_{i=1}^n D_{q_i}L dq_i - \frac{\partial}{\partial t}L dt \\
&= \sum_{i=1}^n (p_i - D_{\dot{q}_i}L) d\dot{q}_i + \sum_{i=1}^n \dot{q}_i dp_i - \sum_{i=1}^n D_{q_i}L dq_i - \frac{\partial}{\partial t}L dt.
\end{aligned}$$

Since $p=D_{\dot{q}}L$, we have $dH=\sum_{i=1}^n \dot{q}_i dp_i - \sum_{i=1}^n D_{q_i}L dq_i - \frac{d}{dt}L dt$. In other words, H is only a function of q,p,t. Then $dH=D_pH dp + D_qH dq + D_tH dt$. Comparing this to the calculation above, we have that $D_qH=-D_qL$.

Then $\dot{p}=-D_qH \iff \dot{p}=D_qL \iff \frac{d}{dt}(D_{\dot{q}}L)=D_qL$, which concludes the proof.

□

Remark 2.16. Notice that here we assume $D_{\dot{q}\dot{q}}L$ to be invertible so that (q,p) can also be considered as local coordinate systems in TM in replacement of $(q,\dot{q})$ by inverse function theorem since then the transformation $(q,p)=(q,D_{\dot{q}}L)$ would

have invertible differentials, which enables us to apply inverse function theorem and get local smooth inverse functions for the transformation map.

Combining proposition 2.15 and Theorem 2.13, we see that an action I w.r.t L is stationary if and only if u obeys the Hamiltonian equations of H obtained from a Legrenge transformation from L.

We end this subsection by remarking that the stationary action principle is the fundamental axiom with proper L when it comes to Hamiltonian mechanics, which is why Hamiltonian equations determine the 'correct' trajectory.

2.3. **Canonical transformation and Liouville's theorem.** In this subsection, we discuss the concept of Canonical transformation, which is closely related to Hamiltonian systems discussed in the previous function. To begin with, we define a specific differential 2-form on a $2n-$dimensional manifold M.

Definition 2.17. Given M a $2n-$dimensional manifold with a coordinate system (q,p), we define the 2-form $\omega : \bigcup_{p\in M} T_pM \times T_pM \to \mathbb{R}$ to be:

$$\omega = \sum_{i=1}^{n} dp_i \wedge dq_i \tag{2.18}$$

and we call it a symplectic form.

A general symplectic form on M is any $2-$form on M that is closed and non-degenerate.

Definition 2.19. Given a diffeomorphism $F : M \to M$, we say that F is a canonical transformation if $F^*\omega = \omega$ where $F^*\omega$ refers to the pullback of ω with respect to F.

Since any canonical transformation is a diffeomorphism, we can take it as a transformation of coordinates: $(q,p) \to (Q,P)$. We will show that a canonical transformation is equivalent to a transformation of coordinates that preserves Hamiltonian equations, i.e., q,p satisfy the Hamiltonian equations generated by H if and only if Q,P also satisfy the Hamiltonian equations generated by H.

Lemma 2.20. *Given a diffeomorphism $F(q,p) := (Q(q,p),P(q,p))$, F is a canonical transformation if and only if for all $1 \le i,j \le n$*

$$\{P_i,P_j\} = 0, \quad \{Q_i,Q_j\} = 0 \quad \{Q_i,P_j\} = \delta_{ij} \tag{2.21}$$

where $\{\cdot\}$ is the Poisson bracket defined as $\{f,g\} = \sum_i \frac{\partial f}{\partial q_i}\frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial g}{\partial q_i}$ when (q,p) is the coordinate system.

Proof. The proof follows immediately from the following calculation:

$$\begin{aligned}(Q,P)^*\omega &= (\frac{\partial Q_j}{\partial q_i}\frac{\partial P_j}{\partial p_i} - \frac{\partial Q_j}{\partial p_i}\frac{\partial P_j}{\partial q_i})dq_j \wedge dp_j \\ &= \{Q_j,P_j\}dq_j \wedge dp_j \\ &= \omega \text{ (iff the transformation is canonical).}\end{aligned}$$

□

Lemma 2.22. *$F(q,p) := (Q(q,p),P(q,p))$ is a canonical transformation if and only if given any $f,g : M \to \mathbb{R}$, we have that*

$$\{f,g\}|_{qp} = \{f,g\}|_{QP} \tag{2.23}$$

Proof.

$$\begin{aligned}
\{f,g\}|_{qp} &= \sum_i \frac{\partial f}{\partial q_i}\frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial g}{\partial q_i} \\
&= \sum_i [(\sum_j \frac{\partial f}{\partial Q_j}\frac{\partial Q_j}{\partial q_i} + \frac{\partial f}{\partial P_j}\frac{\partial P_j}{\partial q_i})(\sum_k \frac{\partial g}{\partial Q_k}\frac{\partial Q_k}{\partial p_i} + \frac{\partial g}{\partial P_k}\frac{\partial P_k}{\partial p_i}) \\
&\quad - (\sum_j \frac{\partial f}{\partial Q_j}\frac{\partial Q_j}{\partial p_i} + \frac{\partial f}{\partial P_j}\frac{\partial P_j}{\partial p_i})(\sum_k \frac{\partial g}{\partial Q_k}\frac{\partial Q_k}{\partial q_i} + \frac{\partial g}{\partial P_k}\frac{\partial P_k}{\partial q_i})] \\
&= \sum_i \sum_j \sum_k \frac{\partial f}{\partial Q_j}\frac{\partial g}{\partial P_k}\{Q_j, P_k\} - \frac{\partial f}{\partial P_j}\frac{\partial g}{\partial Q_k}\{P_j, Q_k\} \\
&\quad + \frac{\partial f}{\partial Q_j}\frac{\partial g}{\partial Q_k}\{Q_j, Q_k\} - \frac{\partial f}{\partial P_j}\frac{\partial g}{\partial P_k}\{P_j, P_k\}.
\end{aligned}$$

On the other hand,

$$\{f,g\}|_{QP} = \sum_i \frac{\partial f}{\partial Q_i}\frac{\partial g}{\partial P_i} - \frac{\partial f}{\partial P_i}\frac{\partial g}{\partial Q_i}. \tag{2.24}$$

Therefore, $\{f,g\}|_{qp} = \{f,g\}|_{QP}$ for any f, g if and only if

$$\{P_i, P_j\} = 0, \quad \{Q_i, Q_j\} = 0 \quad \{Q_i, P_j\} = \delta_{ij}$$

which, by Lemma 2.20 is equivalent to F being a canonical transformation. □

With these lemmas established, we can finally move to the proposition that links canonical transformation and preservation of Hamiltonian.

Proposition 2.25. *$F(q,p) := (Q(q,p), P(q,p))$ is a canonical transformation if and only if whenever $x(t) = (q(t), p(t))$ satisfies the Hamiltonian equations generated by a function H under coordinate system (q,p), it also satisfies the Hamiltonian equations generated by the same function H under the transformed coordinate system $(Q.P)$*

Proof. The key observation is that we can rewrite Hamiltonian equations w.r.t H under coordinate system (q,p) in terms of Poisson brackets:

$$\dot{q} = \{q, H\}_{pq}, \quad \dot{p} = \{p, H\}_{pq}. \tag{2.26}$$

Then for each q, p, H, $\dot{q} = \{q,H\}_{pq}$, $\dot{p} = \{p,H\}_{pq} \iff \dot{Q} = \{Q,H\}_{pq}$, $\dot{P} = \{P,H\}_{pq}$. It then follows that all corresponding Q, P satisfies $\dot{Q} = \{Q,H\}_{PQ}$, $\dot{P} = \{P,H\}_{PQ}$ is equivalent to $\{\cdot\}_{pq} \equiv \{\cdot\}_{PQ}$, which, by Lemma 2.22, is equivalent to $F = (Q,P)$ being a canonical transformation. □

Therefore, canonical transformation can be used to transform the coordinates into such that the corresponding Hamiltonian vector field has better properties. For example, action-angle coordinates make sure half of the coordinates are invariant along the trajectory, whereas Hamilton-Jacobi theory aims at finding Hamiltonian vector fields in which everything is fixed.

We will now revisit the famous result that the flow of any Hamiltonian vector field preserves volume. Additionally, we will generalize the result to all vector fields with 0 divergence. Innovatively, we will show that a proof using Stoke's theorem can

also be used to show the generalization of the result of not only volume preservation but also the flow being canonical transformations to all 0 divergence vector fields.

Theorem 2.27. *For any Hamiltonian H, the flow of the corresponding Hamiltonian vector field at any time t is a canonical transformation.*

Proof. To start off, we consider the extended phase space, i.e., $TM \times \mathbb{R}$ which contains the phase space in addition to time (q, p, t), and we define a one-form on this space:

$$\sigma^1 = pdq - Hdt \tag{2.28}$$

where H is any Hamiltonian. We will now see that there exists a vector V such that $d\sigma^1 \lfloor V = 0$, i.e., $d\sigma^1(V, \eta) = 0$ for any η. The V direction is also called vortex lines. It's easy to check that the following V works:

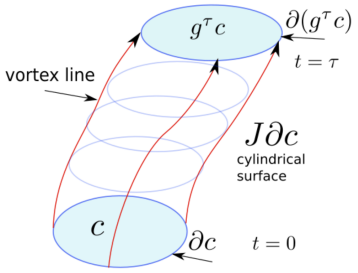
$$V = D_p H \frac{\partial}{\partial q} - D_q H \frac{\partial}{\partial p} + \frac{\partial}{\partial t}.$$

Remark 2.29. Notice that the flow generated by this vector field satisfies the Hamiltonian equations, which means the flows of V are basically just Hamiltonian flows.

We now consider any arbitrary region in the phase space that is an orientable manifold and call it C. And consider its Hamiltonian flow along t from $t = 0$ to $t = \tau$, causing a transformation $g^\tau C$. To make it more precise, consider a function $g : TM \times \mathbb{R} \to TM \times \mathbb{R}$ defined as the flowing:

$$g(x, t) = (F_H^t(x), t),$$

where $x \in M$ and F_H^t is the hamiltonian flow. Then applying this mapping on all the points on C from $t = 0$ to $t = \tau$ should give us a "cylinder", like the following [12]:



Now, we recall the generalized Stoke's theorem:

Theorem 2.30. *Let C be an orientable manifold, and ω be a differential form on that region, we have that*

$$\int_C d\omega = \int_{\partial C} \omega.$$

One consequence is the integration over a boundary of an exact C^2 form is going to be 0. In particular, we have that

$$\int_{\partial C} d\sigma^1 = \int_C d^2\sigma^1 = 0.$$

The curves along it are like the graph of the integral curves of the Hamiltonian field. We can now break the above integration over the whole boundary of the cylinder into three parts. We first notice that $\int_{J\partial C} d\sigma^1 = 0$. To see this, we look at a parameterization $G : K \times [a, b] \to TM \times \mathbb{R}$ of this boundary:

$$(k, t) \to (F_H^t(\lambda(k)), t)$$

where λ is a parameterization of C from $K \subset \mathbb{R}^{2n}$. We then have that

(2.31)
$$\frac{dG}{dt} = (\frac{dF_H^t(\lambda(k))_q}{dt}, \frac{dF_H^t(\lambda(k))_p}{dt}, 1) = (D_pH(\lambda(k)), -D_qH(\lambda(k)), 1) = V(\lambda(k)).$$

Consequently,

$$\int_{J\partial C} d\sigma^1 = \int_{K\times[a,b]} d\sigma^1(D_kG, D_tG) = 0 \tag{2.32}$$

since $\int d\sigma^1(V, ...) = 0$, which gives us the conclusion. (This also gives us a nice property for σ^1, which is that integrating $d\sigma^1$ over a parameterization that involves the corresponding integral curve always gives us 0). Consequently integrating $d\sigma^1$ on the top and both faces of the cylinder C and $g^\tau C$ must give the same result.

Now, integrating over $d\sigma^1$ on these two surfaces is the same as integrating over $\omega = dp \wedge dq$ since all vectors are perpendicular to $\frac{\partial}{\partial t}$. To put everything together, we have that:

$$\int_A \sigma = \int_{F^t(A)} \sigma = \int_A {F^t}^* \sigma$$

for any A in the phase space. This in turn implies that we must have $\sigma = {F^t}^*\sigma$. □

Corollary 2.33. *(Liouville's theorem) For any Hamiltonian H, the flow of the corresponding Hamiltonian vector field at any time t preserves the volume form.*

Proof. Since the symplectic two-form can be wedged product together and gives the volume in phase space, we also have that the flow preserves volume in the phase space since ${F^t}^*(\sigma \wedge ... \wedge \sigma) = {F^t}^*\sigma \wedge ... \wedge {F^t}^*\sigma = \sigma \wedge ... \wedge \sigma$. □

We will now look at another proof of Liouville's theorem that can be generalized to not only flows of Hamiltonian vector fields but also all vector fields with 0 divergence. Notice that any Hamiltonian vector field has 0 divergence:

$$divH_H = \frac{\partial}{\partial p}(-\frac{\partial H}{\partial q}) + \frac{\partial}{\partial q}(-\frac{\partial H}{\partial p}) \equiv 0.$$

Inspired by this result, we then make an observation undocumented before: the flow of any vector field with 0 divergence on a contractible manifold doesn't only preserve volume, but is also canonical, using a generalization of the proof using Stoke's theorem. To begin with, we first establish the following theorem:

Theorem 2.34. *Define $V(t) = \int_{F^t(D)} dx$ where D is a bounded, measurable region in the phase space and F^t being the flow for a vector field $f : \mathbb{R}^n \to \mathbb{R}^n$. Then we have that*

$$\dot{V}(t) = \int_{F^t(D)} div(f(x))dx.$$

Remark 2.35. Notice that this directly gives Liouville's theorem as a corollary since $divH_H = 0$.

Now, we dive into the proof of this theorem:

Proof. Notice that $V(t_0+h) = \int_{F^{t_0+h}(D)} dy = \int_{F^h(F^{t_0}(D))} dy = \int_{F^{t_0}(D)} J(F^h(x))dx$. And it suffices to show that $J(F^h(x)) = 1 + h \cdot div(f(x)) + o(h)$.

To see this, we first notice that $F^h(x) = x + \int_0^h f(F^s(x))ds$. Then we consequently have that $DF^h(x) = I + \int_0^h Df(F^s(x))ds$ by Leibniz rule. As a consequence, this gives us $\frac{\partial DF^h(x)}{\partial h} = Df(F^h(x))$. By definition of the derivative, this gives us

$$DF^h(x) = I + hDf(x) + o(h). \tag{2.36}$$

This consequently gives us the result we want since

$$\begin{aligned} J(F^h(x)) &= det(I + hDf(x)) + o(h) \text{ by continuity of det} \\ &= \Pi_{i=1}^n (1 + h\frac{\partial f_i}{x_i}(x)) + o(h) \\ &= 1 + h\sum_{i=1}^n \frac{\partial f_i}{x_i}(x) + o(h). \end{aligned}$$

□

Remark 2.37. I want to briefly discuss the intuition for (2.36) above. Given small h, the left-hand side corresponds to the instantaneous change of displacement along integral curves in any direction. The right-hand side is roughly the change of instantaneous velocity times time h along the integral curves by any direction, plus the difference due to the slight change of x itself. Intuitively, these should be about the same value.

Finally, with light shed on the generalization of the conclusion, we conclude this section by seeing how the proof involving Stokes theorem can generalize to all vector fields with 0 divergence on a contractible manifold.

Theorem 2.38. *For any vector field V on a contractible manifold with 0 divergence, the flow of V at any time t is a canonical transformation*

Proof. In Stokes' theorem proof, it involves integrating over a form $d\sigma^1 = dp \wedge dq - \frac{\partial H}{\partial q} dq \wedge dt - \frac{\partial H}{\partial p} dp \wedge dt$. And then showing that $d\sigma^1(V, \cdot) \equiv 0$ where $V = (\frac{\partial H}{\partial q}, -\frac{\partial H}{\partial q}, 1)^T$.

However, notice that for any vector field V, if we talk about the flow induced by it and construct the same cylinder using that flow, we will get a parameterization involving $\frac{dG}{dt} = (V_q, V_p, 1)^T$. And if we consider integrating the boundary of the cylinder over a similar form $\omega' = dp \wedge dq - (-V_p)dq \wedge dt - V_q dp \wedge dt$, we can get that $\int_{J\partial c} \omega' = 0$. Therefore, we have that the flow of V preserves volume if ω' is exact (by applying Stoke's theorem as before).

The crucial observation is that, ω' is exact iff V has 0 divergence:

Consider $d\omega$, we have that

$$\begin{aligned} d\omega' &= -(-\frac{\partial V_p}{\partial p})dp \wedge dq \wedge dt - \frac{\partial V_q}{\partial q} dp \wedge dq \wedge dt \\ &= div(V)dp \wedge dq \wedge dt. \end{aligned}$$

And by Poincare's lemma, since the domain is contractible, we have that $d\omega' = 0 \implies \omega'$ is exact. Therefore, $div(V) = 0 \implies \omega'$ is exact. Conversely, if ω' is exact, $d\omega' = 0$, which means that $div(V) = 0$. □

3. Sherman's Area convexity Algorithm and symplectic conjugate as Hamiltonian

3.1. **Sherman's Area convexity Algorithm.** In this section, we will analyze the area convexity algorithm proposed in Sherman's paper. Sherman used this algorithm to break the $l_\infty-$barrier and solved right-stochastic matrix problems, and consequently multicommodity flow problem in accelerated time. However, this algorithm is somewhat ad-hoc and requires more examination. One potential direction is to interpret this algorithm in terms of Hamiltonians, which is not made clear in any literature. Therefore, the goal of this section is to make some connections between this algorithm and Hamiltonian systems and symplectic structures and shed some light on this direction.

Before linking the algorithm with symplectic structures and Hamiltonian systems, we will first set up the problem that the algorithm is targeting, and then introduce the algorithm formally.

We consider the general bi-affine saddle point problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\langle y, Ax\rangle - \langle b, y\rangle - \langle c, x\rangle \tag{3.1}$$

where X, Y are compact-convex in real finite-dimensional vector spaces, and A is a linear operator from $\mathcal{X}$ to Y^*, b, c are linear functionals. We will reduce this problem to a purely bilinear, self-dual form. By increasing the dimension by 1 and augmenting $\mathcal{X}, \mathcal{Y}$ with constant coordinates, we can turn the above problem into the purely bilinear form:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\langle y, Ax\rangle. \tag{3.2}$$

Notice that when the duality gap is 0, i.e.

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\langle y, Ax\rangle - \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\langle y, Ax\rangle = 0,$$

the primal points (x, y) is a solution to the saddle point problem (3.2). This is because each $\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\langle y, Ax\rangle$ is a lower bound to $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\langle y, Ax\rangle$. By the same reason, if we have $(x, y), (x', y')$ that make the duality gap only $\delta > 0$, then $\langle y, Ax\rangle$ is at most δ away from $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\langle y, Ax\rangle$. This is what we call a $\delta-$**approximate solution** for (3.2).

We now transform the duality gap as the following:

$$\begin{aligned}
&\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\langle y, Ax\rangle - \max_{y'\in\mathcal{Y}}\min_{x'\in\mathcal{X}}\langle y', Ax'\rangle \\
=&\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\langle y, Ax\rangle - (-\min_{y'\in\mathcal{Y}}\max_{x'\in\mathcal{X}}\langle y', -Ax'\rangle) \\
=&\min_{x\in\mathcal{X}, y'\in\mathcal{Y}}\max_{x'\in\mathcal{X}, y\in\mathcal{Y}}\langle y, Ax\rangle - \langle y', Ax'\rangle.
\end{aligned}$$

Now, let $\mathcal{C} = \mathcal{X}\oplus\mathcal{Y}$ and let J be the linear operator on $\mathcal{C}$ to $\mathcal{C}^*$ such that $J(x, y) = (A^*y, -Ax)$. Notice that $J^* = -J$, i.e., J is alternating. To see this, we have that

$$\begin{aligned}
\langle (x, y), J^*(x', y')\rangle &= \langle (x, y), -J(x', y')\rangle \\
&= \langle (x, -A^*y')\rangle + \langle (y, Ax')\rangle \\
&= \langle x', A^*y\rangle + \langle y', -Ax\rangle \\
&= \langle (x', y'), J(x, y)\rangle.
\end{aligned}$$

Now, we can express the above duality gap in terms of J:

$$\begin{aligned}
&\min_{x\in\mathcal{X},y'\in\mathcal{Y}} \max_{x'\in\mathcal{X},y\in\mathcal{Y}} \langle y, Ax\rangle - \langle y', Ax'\rangle \\
=& - \min_{x\in\mathcal{X},y'\in\mathcal{Y}} \max_{x'\in\mathcal{X},y\in\mathcal{Y}} \langle x', A^*y'\rangle - \langle y, Ax\rangle \\
=& - \min_{x\in\mathcal{X},y'\in\mathcal{Y}} \max_{x'\in\mathcal{X},y\in\mathcal{Y}} \langle (x', y), (A^*y', -Ax)\rangle \\
=& - \min_{x\in\mathcal{X},y'\in\mathcal{Y}} \max_{x'\in\mathcal{X},y\in\mathcal{Y}} \langle (x', y), J(x, y')\rangle \\
=& - \min_{z\in C} \max_{z'\in C} \langle z', Jz\rangle.
\end{aligned}$$

Therefore, the problem of minimizing the duality gap reduces to the problem of

$$\min_{z\in C} \max_{z'\in C} \langle z', Jz\rangle. \tag{3.3}$$

Additionally, if the original purely bilinear saddle point problem (3.2) has a solution, then $\min_{z\in C} \max_{z'\in C} \langle z', Jz\rangle = 0$. Therefore, we've turned the bi-affine saddle point problem into a self-dual saddle point problem with value 0.

We now move to present Sherman's algorithm to solve this self-dual problem. Before that, we list some required definitions:

Definition 3.4. Given $\phi : \mathcal{C} \to \mathbb{R}$ where $\mathcal{C} \subset C$, we define the $\delta-$approximate minimization oracle ($\delta-$AMO) to be a map $\Phi : \mathcal{C}^* \to \mathcal{C}$ such that $\Phi(a)$ satisfies

$$\langle a, \Phi(a)\rangle - \phi(\Phi(a)) + \delta \geq \phi^*(a) \tag{3.5}$$

where ϕ^* is the Fenchel-conjugate of ϕ.

Definition 3.6. We say that a map $\phi : \mathcal{C} \to \mathbb{R}$ where $\mathcal{C}$ is convex to be $\eta-$area-convex with respect to J on convex set $\mathcal{C}$ iff for all $x, y, z \in \mathcal{C}$

$$\eta\langle y - z, J(y - x)\rangle \leq \phi(x) + \phi(y) + \phi(z) - 3\phi(\frac{x + y + z}{3}). \tag{3.7}$$

We are now ready to present the following algorithm:

Algorithm 1 `Find_Saddle_point`

Input: $\phi : \mathcal{C} \to [-\rho, 0]$, $\mathcal{C}$ compact-convex in C, and ϕ (-2)-area-convex with respect to J; Φ being a $\delta-$AMO for ϕ; $\epsilon > 0$
Output: a $\delta + \epsilon$ approximate solution for self-dual saddle point problem (3.3) over $\mathcal{C}$
$z(0) \leftarrow 0$
for $0 \leq t < \rho\epsilon^{-1}$ **do**
    $z(t + 1) = z(t) + \Phi(Jz(t) + 2J\Phi(Jz(t)))$
end for
Return $\frac{z(\lfloor\rho\epsilon^{-1}\rfloor+1)}{\lfloor\rho\epsilon^{-1}\rfloor+1}$

Remark 3.8. Notice that we assumed 0 to be in $\mathcal{C}$ since $z(0) = 0$, and this implies that for all $t \in \mathbb{N}_{>0}$, we have that $\frac{z(t)}{t} \in \mathcal{C}$ by convexity of $\mathcal{C}$. And this justifies how the returned value is a valid candidate for a $\delta+\epsilon$ approximate solution for self-dual saddle point problem (3.3) over $\mathcal{C}$.

Theorem 3.9. *Algorithm 1 outputs a $\delta+\epsilon$ approximate solution for self-dual saddle point problem (3.3)*

Proof. Since the value of problem (3.3) is 0, it suffices to prove that when $t = \rho\epsilon^{-1}$, $\langle z', \frac{z(t)}{t}\rangle \leq \epsilon + \delta$ for any $z' \in C^*$. we can define $\gamma(a) := sup_{z\in\mathcal{C}}\langle a, z\rangle$. And since ϕ is non-positive, $\phi^*(a) = sup_{z\in\mathcal{C}}\langle a, z\rangle - \phi(z) \geq \gamma(a)$. Therefore, it suffices to bound ϕ^* above by $\delta + \epsilon$.

And one crucial Lemma to that end is that the update scheme in Algorithm 1 guarantees that ϕ^* only deviate by δ each step:

Lemma 3.10. *For any $a \in C^*$, we have that*

$$\phi^*(a + J\Phi(a + 2J\Phi(a)) \leq \phi^*(a) + \delta. \tag{3.11}$$

Let's first check how this lemma gives us the theorem. We have that for each t,

$$\phi^*(Jz(t+1)) = \phi^*(Jz(t) + J\Phi(Jz(t) + 2J\Phi(Jz(t))) \leq \phi^*(Jz(t)) + \delta. \tag{3.12}$$

Therefore, by induction, we can conclude that $\phi^*(Jz(t)) \leq \rho + \delta t$. Notice that $\gamma(\frac{a}{t}) = \frac{\gamma(a)}{t}$. Therefore, $\gamma(\frac{Jz(t)}{t}) = \frac{\gamma(a)}{t} \leq \frac{\phi^*(Jz(t))}{t} \leq \frac{\rho}{t} + \delta$. Consequently, when $t = \rho\epsilon^{-1}$, we have $\gamma(\frac{Jz(t)}{t}) \leq \delta + \epsilon$, which concludes the proof.

Now, we prove Lemma 3.10:

Proof. Assume without loss of generality that $a = 0$ since we can always change $\phi(x)$ into $\phi(x) - \langle a, x\rangle$. Let $x = \Phi(0)$ and $y = \Phi(0 + 2J\Phi(0)) = \Phi(2Jx)$. We want to show that for any $z \in \mathcal{C}$

$$\langle z, Jy\rangle - \phi(z) \leq \phi^*(0) + \delta. \tag{3.13}$$

Now, for $x, y, z \in \mathcal{C}$, since ϕ is (-2)-area convex with respect to J, we have that

(3.14)
$$\begin{aligned}
\langle z - y, J(y - x)\rangle &\leq \frac{1}{2}(\phi(x) + \phi(y) + \phi(z) - \phi(\frac{x + y + z}{3})) \\
&\leq \frac{1}{2}(\phi(x) + \phi(y) + \phi(z) + 3\phi^*(0)) \text{ (by definition of conjugate)} \\
&\leq \frac{1}{2}(\delta + \phi(y) + \phi(z)) + \phi^*(0) \\
&\text{(since } x = \Phi(0), \text{ we have } -\phi(x) \geq \phi^*(0) - \delta).
\end{aligned}$$

Now, since $y = \Phi(2Jx)$, we have $\langle y, 2Jx\rangle - \phi(y) + \delta \geq \langle z, 2Jx\rangle - \phi(z)$, which is equivalent to

$$\langle z - y, Jx\rangle - \phi(z) \leq \frac{1}{2}(\delta - \phi(y) - \phi(z)). \tag{3.15}$$

Combining this with (3.14), we finally have that

$$\begin{aligned}
\langle z, Jy\rangle - \phi(z) &= \langle z - y, Jy\rangle - \phi(z) \\
&= \langle z - y, J(y - x)\rangle + \langle z - y, Jx\rangle - \phi(z) \\
&\leq \delta + \phi^*(0).
\end{aligned}$$

□

With the proof of Lemma 3.10, we conclude the proof of Theorem 3.9 □

3.2. **symplectic conjugate and Hamiltonian perspective.** In the rest of this paper, we will tackle two things about this algorithm. Firstly, the update method chosen in this algorithm seems somewhat mysterious, we will try to introduce a symplectic and Hamiltonian perspective to the behavior of the algorithm in this section. Secondly, it's still unclear how is this algorithm precisely linked to more general and classical optimization schemes. In Section 4, we will establish a clear relationship between Algorithm 1 and a variation of dual-extrapolation.

The intuition behind linking Algorithm 1 with symplectic structures is the close relationship between symplectic form and saddle point problems. We've already seen that every bi-affine saddle point problem can be reduced to a self-dual saddle point problem (3.3) in which J is alternating. However, notice that there is a nice correspondence between alternating operators and general symplectic forms, i.e., non-degenerate and closed:

Proposition 3.16. *For every general symplectic form on finite-dimensional self-dual C, there exists a unique bijective alternating linear operator J such that*

$$\omega(x, y) = \langle y, Jx\rangle \tag{3.17}$$

for all $x.y \in C$. Conversely, for every bijective alternating linear operator J, there exists a unique general form that achieves (3.17).

Proof. Consider the map $J \mapsto (x, y \mapsto \langle y, Jx\rangle)$ from bijective linear operators to general symplectic forms (Notice $(x, y \mapsto \langle y, Jx\rangle)$ is a general symplectic form whenever J is bijective and alternating). Now, it suffices to show that this map is bijective.

It's injective since symplectic forms are non-degenerate. To see that it's surjective, for each ω, let J be $Jx = (y \mapsto \omega(x, y))$, then (3.17) is satisfied. Finally, we need to make sure J here is bijective. This can be seen from the fact that ω is non-degenerate, and $dim\, C = dim\, C^*$ are finite. □

In close examination of the proof of Theorem 3.9, we realize that the control of the value of $\phi^*(Jz)$ is crucial for the guarantee of accelerated convergence rate. Therefore, this relationship we've established between symplectic form and bijective, alternating J inspires us to construct a new type of conjugate using symplectic forms for ϕ that encodes this information of alternating linear operator naturally.

Definition 3.18. Given $\phi : C \to \mathbb{R}$, let $\phi^{*\omega} : C \to \mathbb{R}$ be that

$$\phi^{*\omega}(y) = \sup_{x \in C} \omega(y, x) - \phi(x). \tag{3.19}$$

This obviously satisfies an inequality analogous to Fenchel inequality for all $x, y \in C$:

$$\phi(x) + \phi^{*\omega}(y) \geq \omega(y, x). \tag{3.20}$$

At the same time, we define the symplectic subgradient for a function in analogy with the normal subgradient:

Definition 3.21. Given $\phi : C \to \mathbb{R}$, we define:

$$y \in \partial^\omega \phi(x) \iff \phi(z) \geq \phi(x) + \omega(y, z - x) \text{ for all } z \in C. \tag{3.22}$$

Now, we prove that a function takes exactly those that make the symplectic Fenchel inequality tight in its symplectic subgradient, i.e.,

Theorem 3.23. $y \in \partial^\omega \phi(x) \iff \phi(x) + \phi^{*\omega}(y) = \omega(y, x)$.

Proof. We only need to prove $y \in \partial^\omega \phi(x) \iff \phi(x) + \phi^{*\omega}(y) \leq \omega(y, x)$.

By definition, $y \in \partial^\omega \phi(x) \iff \phi(z) \geq \phi(x) + \omega(y, z - x)$ for all $z \in C$, The latter is equivalent to $\omega(y, x) - \phi(x) \geq \omega(y, z) - \phi(z)$ for all $z \in C$, which is the same as $\omega(y, x) - \phi(x) \geq \phi^{*\omega}(y)$. This concludes the proof. □

Before doing more analysis on symplectic conjugate and subgradient, let's try to build the relationship between them and the ordinary conjugate and subgradient and see how the symplectic conjugate encodes the information of an alternating linear operator into Fenchel's conjugate. It's easy to check the following:

$$y \in \partial^\omega \phi(x) \iff J(y) \in \partial \phi(x) \tag{3.24}$$

$$\phi^{*\omega}(x) = \phi^*(J(x)) \tag{3.25}$$

$$\Phi(Jz) = \Phi^\omega(z) \text{ where } \Phi \text{ represents the } 0-\text{AMO} \tag{3.26}$$

where J is the corresponding linear operator for ω as proposed in Proposition 3.16. Since $d(\phi^*)(x) = \Phi(x)$, we consequently have that $d(\phi^{*\omega})(x) = \Phi^\omega(x)J$ by the chain rule.

We now move to an important result that is analogous to the conjugate correspondence between smoothness and strong convexity for Fenchel conjugate. We similarly have that area-convexity is equivalent to the local smoothness of the symplectic conjugate in terms of the corresponding symplectic form for a twice differentiable ϕ. This doesn't only shed some light on the nature of area-convexity but also builds up the reasonability to consider symplectic conjugate as a reasonable notion of conjugate.

We first put up a Lemma proposed in Sherman's paper and show some other preliminary lemmas.

Lemma 3.27. *[Sherman 17] Let ϕ be twice differentiable on convex C,*

(a) If ϕ is $(-\frac{1}{\sqrt{3}})-$area convex with respect to J on the interior of C, then $d^2\phi(z) \succeq \iota J$ in the interior of C

(b) If $d^2\phi(z) \succeq \iota J$ for all $z \in C$, then ϕ is $-\frac{1}{3\sqrt{3}}$-area convex with respect to J on C

where we define $Q \succeq \iota J$ to be

$$\begin{pmatrix} Q & -J \\ J & Q \end{pmatrix} \succeq 0. \tag{3.28}$$

Lemma 3.29. *When Q is a positive semi-definite and invertible matrix and alternating matrix J, we have that $Q \succeq \iota J$ if and only if $Q \succeq J^*QJ$*

Proof. We notice that

$$\begin{pmatrix} I & 0 \\ -J^*Q^{-1} & I \end{pmatrix} \begin{pmatrix} Q & J^* \\ J & Q \end{pmatrix} \begin{pmatrix} I & -Q^{-1}J \\ 0 & I \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & Q - J^*Q^{-1}J \end{pmatrix}.$$

Since the left and the right matrix on the left side of the equation are all invertible, $\begin{pmatrix} Q & J^* \\ J & Q \end{pmatrix}$ is positive semi-definite if and only if $\begin{pmatrix} Q & 0 \\ 0 & Q - J^*Q^{-1}J \end{pmatrix}$ is positive semi-definite, which is equivalent to $Q - J^*Q^{-1}J \succeq 0$ since $Q \succeq 0$, which concludes the proof. □

Lemma 3.30. *Given twice-differentiable function ϕ with invertible Hessian and an alternating operator J, we have that $(d^2(\phi(z)))^{-1}J$ is a contraction if and only if $d^2(\phi(z)) \succeq J^*(d^2(\phi(z)))^{-1}J$*

Theorem 3.31. *Given $\phi : C \to \mathbb{R}$ to be twice-differentiable and Φ is bijective (which is the case when ϕ strictly convex), also assume J to be invertible:*

(a) If ϕ is $(-\frac{1}{\sqrt{3}})-$area-convex w.r.t J , then we have that Φ^ω is locally $((1+\epsilon)$ for any $\epsilon > 0)$ Lipschitz.

(b) If Φ^ω is locally $L-$Lipschitz, then ϕ is $-\frac{1}{3\sqrt{3}L}-$area-convex w.r.t J.

Before going into the proof, I want to illustrate the analogy between this theorem and the conjugate correspondence of smoothness and strong convexity. Φ^ω being Lipschitz is not exactly the same as $\phi^{*\omega}$ being smooth since $d\phi^{*\omega}(x) = \Phi^\omega(x)J$. It's stronger than regular smoothness since

$$||\Phi^\omega(z_0) - \Phi^\omega(z)|| \leq L||z_0 - z|| \implies ||J(\Phi^\omega(z_0) - \Phi^\omega(z))|| \leq ||J||L||z_0 - z||.$$

We now proof the above theorem:

Proof. (a) If ϕ is strictly $-\frac{1}{\sqrt{3}}-$area-convex w.r.t J, by Lemma 3.27, we have that $d^2\phi(z') \succeq \iota J$ for all $z' \in C$. This is equivalent to $J^*(d^2\phi(z'))^{-1}J \preceq d^2\phi(z')$ by Lemma 3.29. Notice that we can apply Lemma 3.29 here since the area-convexity of ϕ guarantees the Hessian to be positive semi-definite.

Now, $J^*(d^2\phi(z'))^{-1}J \preceq d^2\phi(z')$ in turn is equivalent to $(d^2\phi(z'))^{-1}J$ being a contraction according to Lemma 3.30.

By second-order duality theory, we can show that $d^2\phi^*(Jz) = (d^2\phi(z'))^{-1}$ where $z' = \Phi(Jz)$. Notice that this is always possible since Φ is bijective. Along with the fact that J is invertible, this implies $d^2\phi^*(Jz)J = d\Phi(Jz)J = d(\Phi \circ J)(z) = d\Phi^\omega(z) = (d^2\phi(z'))^{-1}J$.

For any $z_0 \in C$, by Taylor's theorem, we have that

$$\Phi^\omega(z_0) = \Phi^\omega(z) + d\Phi^\omega(z)(z_0 - z) + \frac{1}{2}\langle(z - z_0), h_2(z_0)(z - z_0)\rangle$$

where $h_2(z_0) \to 0$ as $z_0 \to z$. Therefore, we have that

$$\begin{aligned}
||\Phi^\omega(z_0) - \Phi^\omega(z)|| &= ||d\Phi^\omega(z)(z_0 - z) + \frac{1}{2}\langle(z - z_0), h_2(z_0)(z - z_0)\rangle|| \\
&\leq ||d\Phi^\omega(z)(z_0 - z)|| + \frac{1}{2}||h_2(z_0)(z - z_0)||||z - z_0|| \\
&< ||z_0 - z|| + \frac{1}{2}||h_2(z_0)(z - z_0)||||z - z_0|| \text{ since } d\Phi^\omega(z) \text{ is a contraction.}
\end{aligned}$$

Notice, if we pick z_0 close enough to z, we have that $\frac{||h_2(z_0)(z-z_0)||||z-z_0||}{2k||z_0-z||}$ goes to 0 for any $k > 0$, which implies that we can pick a neighborhood of z such that $\frac{1}{2}||h_2(z_0)(z-z_0)||||z-z_0|| \leq k||z_0-z||$, which shows that Φ^ω is locally Lipschitz at z. Since Φ is bijective, this shows that Φ^ω is locally $(1+\epsilon)$-Lipschitz for any $\epsilon > 0$.

(b) Conversely, if Φ^ω is locally $L-$Lipschitz. Then for any $z \in C$, there exists δ such that for any z_0 such that $||z_0 - z|| \leq \delta$, we have

$$||\Phi^\omega(z_0) - \Phi^\omega(z)|| \leq L||z_0 - z||.$$

By the Taylor expansion, we have that

$$||d\Phi^\omega(z)(z_0 - z)|| - \frac{1}{2}||h_2(z_0)(z - z_0)||||z - z_0|| \leq L||z_0 - z||$$

and if we pick δ' to be small enough, it follows that

$$||d\Phi^{\omega}(z)(z_0 - z)|| \leq L||z_0 - z|| \text{ for all } z_0 \in N_{\delta'}(z).$$

In other words,

$$\frac{1}{L}d(\Phi \circ J)(z) = d(\Phi \circ \frac{J}{L})(z) \text{ is a contraction in the } \delta' \text{ neighborhood of } z$$

But since $d(\Phi \circ \frac{J}{L})(z)$ is linear, this means it's a global contraction. And by what's established above, this is equivalent to $d^2\phi(z') \succeq \iota\frac{J}{L}$ where $z' = \Phi(Jz)$. Notice that $(\Phi \circ J)$ is bijective since both Φ and J are, which implies $d^2\phi(z') \succeq \iota\frac{J}{L}$ on all $z' \in C$. By Lemma 3.27, we finally conclude that ϕ is $-\frac{1}{3\sqrt{3}L}-$area-convex w.r.t J.

□

With this symplectic conjugate being established, we make the key observation that we can view Algorithm 1 as a variational way of following the integral curve of a particular Hamiltonian field with a controlled deviation from the integral curve in terms of energy.

Consider a naive version of the update step in Algorithm 1 when Φ is the $0-$AMO:

$$z(t+1) = z(t) + \Phi(Jz(t)). \tag{3.32}$$

This is the normal explicit Euler discretization of the following continuous dynamic system:

$$\frac{dz}{dt} = \Phi(Jz(t)) = \Phi^{\omega}(z(t)), \quad z(0) = 0. \tag{3.33}$$

This points us to the direction of taking the Hamiltonian to be $H(z) = \phi^{*\omega}(z)$. Since we do have that $\frac{dH(z)}{dz} = \frac{d\phi^{*\omega}(z)}{z} = \frac{d\phi^*(J(z))}{dz} = \Phi(J(z))J = \Phi^{\omega}(z(t))J$, (3.33) can be re-written as:

$$\frac{dz}{dt}J = \frac{dH(z)}{dz}. \tag{3.34}$$

But since J is alternating, we have that $(\frac{dz}{dt}J)^T = -J(\frac{dz}{dt})^T$. Therefore, if we consider $\frac{dz}{dt}$ and $\frac{dH(z)}{dz}$ as column vectors, the above is the same as:

$$-J(\frac{dz}{dt}) = \frac{dH(z)}{dz} \tag{3.35}$$

which is a generalized form of Hamiltonian equations. Consider the case in which $J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$, then $J^{-1} = -J$. So we recover the typical Hamiltionian equations: $(\frac{dz}{dt}) = J\frac{dH(z)}{dz} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\frac{dH(z)}{dz}$. And we can see that energy conservation holds for the more general form of Hamiltonian equations as well:

$$\frac{dH(z(t))}{dt} = \frac{dH(z)}{dz}\frac{dz}{dt} = \frac{dz}{dt}J\frac{dz}{dt} = 0.$$

Notice that in the proof of Theorem 3.9, guaranteeing (3.12) suffices to give us the final conclusion. However, notice that this is equivalent to ensuring that Hamiltonian defined as symplectic conjugate doesn't increase up to a δ. And this is ensured when this Hamiltonian is conserved.

In other words, if we have access to the solution of the above generalized Hamiltonian system (3.35), then following this solution would be an update step that grants the same convergence rate in Theorem 3.9. Let's rewrite the update step in Algorithm 1 in terms of symplectic conjugate:

$$z(t+1) = z(t) + \Phi^\omega(z(t) + 2\Phi^\omega(z(t))). \tag{3.36}$$

With what we've established, we can think of the position at which we access the symplectic conjugate oracle at each step: $z(t) + 2\Phi^\omega(z(t))$ to be established this way to make sure although we might not be able to strictly conserve Φ^ω, we can ensure that it doesn't increase.

With the observation that the guarantee of convergence rate is a result of the nice energy conservation behavior of the update scheme, we end this section by proving that the update scheme (3.36) of Algorithm 1 is not a type of scheme which is well-known to be having nice long-time energy conservation behavior (introducing only a constant deviation of energy): symplectic methods, i.e., numerical methods with each step of update being a canonical transformation.

We first define a type of numerical method called the Runge-Kutta methods and make the observation that the method in Algorithm 1 falls into this category:

Definition 3.37. Let a_{ij}, b_i $(i, j = 1, ..., s)$ be real numbers and let $c_i = \sum_{j=1}^{s} a_{ij}$. An $s-$stage Runge-Kutta method on y is given by:

$$\begin{aligned} k_i &= f(t_0 + c_i h, y_0 + h\sum_{j=1}^{s} a_{ij} k_j) \\ y_1 &= y_0 + h\sum_{i=1}^{s} b_i k_i. \end{aligned} \tag{3.38}$$

If we let $f = \Phi^\omega = \Phi \circ J$, and let $a_{1j} = 0$ for any j which gives us $k_1 = \Phi^\omega(z(t))$. And let $h = 1$, $a_{21} = 2$, we have that $k_2 = \Phi^\omega(z(t) + 2k_1) = \Phi^\omega(z(t) + 2\Phi^\omega(z(t)))$. Finally, we let $b_i = 1$, we recover $z(t+1) = z(t) + k_1 = z(t) + \Phi^\omega(z(t) + 2\Phi^\omega(z(t)))$ which is precisely (3.36). Therefore, this update scheme is indeed a $2-$stage Runge-Kutta method with $a_{11} = a_{12} = 0$, $a_{21} = 2$, $a_{22} = 0$, and $b_1 = b_2 = 1$.

We will now list some results regarding the Runge-Kutta methods:

Definition 3.39. An $s-$stage Runge-Kutta method is called irreducible if for the set of all trees T, we have that the $s \times \infty$ matrix M has full rank s where $M_{i\tau} = g_i(\tau)$ and $g_i(\tau)$ as defined in Hairer $III1.13$

Theorem 3.40. *An irreducible $s-$stage Runge-Kutta method is symplectic if and only if*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \text{ for all } i, j = 1, ..., s. \tag{3.41}$$

Notice that there exists a tree τ with order 3 such that $g_i(\tau) = \sum_{j,k} a_{ij} a_{ik}$. In our case, we have that $g_1(\tau) = 0$ but $g_2(\tau) = 4$. This implies that the two rows of M are linearly independent, which implies that the method in Algorithm 1 is irreducible.

Along twith the fact that for $i = 2, j = 1$, we have

$$b_2 a_{21} + b_1 a_{12} = 2 + 0 = 2 \neq 1 = b_1 b_2. \tag{3.42}$$

Therefore, we conclude that (3.36) is not symplectic.

4. Area Convexity Algorithm and Dual extrapolation

The goal of this section is to target the ad-hoc nature of the discussion concerning Algorithm 1. It's unclear how Algorithm 1 links to more popular and typical optimization methods. In this section, we will close this gap precisely between Algorithm 1 and the method of dual extrapolation. We will also see the correspondence of the Hamiltonian perspective for a special case of dual extrapolation.

4.1. Linking Sherman's Algorithm with dual extrapolation.

Recall that Algorithm 1 requires a $\delta-$AMO Φ for ϕ, and performs the following update:

$$z(t+1) = z(t) + \Phi(Jz(t) + 2J\Phi(Jz(t))). \tag{4.1}$$

On the other hand, the dual extrapolation step proposed by Nesterov is of the following form:

$$(x, y, s_+) \iff \begin{cases} x = T_\beta(\overline{x}, s) \\ y = T_\beta(x, -\lambda g(x)) \\ s_+ = s - \lambda g(y) \end{cases} \tag{4.2}$$

where

$$T_\beta(z, s) = arg\max_x\{\langle s, x - z\rangle - \beta\omega(z, x)\} \tag{4.3}$$

and ω is the Bregman distance, β, λ are coefficients.

The goal of this section is to close the gap between these two methods and shed some light on how (4.1) closely follows (4.2), which is not documented anywhere. To do this, we let $\overline{x}$ be such that $\nabla\phi(\overline{x}) = \Phi(\overline{x}) = 0$ (assuming this is possible) where Φ is the $0-$AMO and let $\beta = 1$, $\lambda = -2$, g to be J, and implicitly we've chosen d (this is involved in the definition of ω) to be ϕ.

Firstly, we claim the following lemma:

Lemma 4.4. *If we choose $\beta = 1$, we have that*

$$T_\beta(z, s) = \Phi(s + \nabla\phi(z)). \tag{4.5}$$

Proof.

$$\begin{aligned} T_\beta(z, s) &= arg\max_x\{\langle s, x - z\rangle - \omega(z, x)\} \\ &= arg\max_x\{\langle s, x - z\rangle - \phi(x) + \phi(z) + \langle\nabla\phi(z), x - z\rangle\} \\ &= arg\max_x\{\langle s + \nabla\phi(z), x - z\rangle - \phi(x) + \phi(z)\} \\ &= arg\max_x\{\langle s + \nabla\phi(z), x\rangle - \phi(x)\} \\ &= \Phi(s + \nabla\phi(z)). \end{aligned}$$

□

Therefore, as an immediate result, we have that $x = \Phi(s + \nabla\phi(\overline{x})) = \Phi(s)$ and $y = \Phi(-\lambda g(x) + \nabla\phi(x)) = \Phi(2Jx + \nabla\phi(x))$. Now, we let $s = Jz(t)$, which leads us to the following:

$$s_+ = Jz_t + 2J\Phi(2J\Phi(Jz_t) + \nabla\phi(\Phi(Jz_t)). \tag{4.6}$$

We now take a closer look at $\nabla\phi(\Phi(Jz_t)$:

Lemma 4.7.

$$\nabla\phi(\Phi(Jz_t) = Jz_t.$$

Proof. We begin by noticing that

$$\phi^*(Jz_t) = \langle Jz_t, \Phi(Jz_t)\rangle - \phi(\Phi(Jz_t)). \tag{4.8}$$

By Danskin's theorem, we have that $\nabla\phi^*(Jz_t) = \Phi(Jz_t)^T J$. On the other hand, from (4.8), we have that

$$\nabla\phi^*(Jz_t) = (\Phi^T(Jz_t) + J^T z_t^T D\Phi(Jz_t))J - \nabla\phi(\Phi(Jz_t))^T D\Phi(Jz_t)J.$$

Therefore, we can conclude

$$\Phi(Jz_t)^T J = (\Phi^T(Jz_t) + J^T z_t^T D\Phi(Jz_t))J - \nabla\phi(\Phi(Jz_t))^T D\Phi(Jz_t)J.$$

If we assume J to be invertible, we can have

$$\begin{aligned}
\Phi(Jz_t)^T &= \Phi^T(Jz_t) + z_t^T J^T D\Phi(Jz_t) - \nabla\phi(\Phi(Jz_t))^T D\Phi(Jz_t)\\
\nabla\phi(\Phi(Jz_t))^T D\Phi(Jz_t) &= z_t^T J^T D\Phi(Jz_t)\\
\nabla\phi(\Phi(Jz_t))^T &= z_t^T J^T \text{ (assuming } D\Phi(Jz_t)) \text{ is invertible).}
\end{aligned}$$

□

All in all, we have that

$$S_+ = Jz_t + 2J\Phi(2J\Phi(Jz_t) + Jz_t). \tag{4.9}$$

If we let $S_+ = Jz_{t+1}$, the above is equivalent to

$$z(t+1) = z(t) + 2\Phi(Jz(t) + 2J\Phi(Jz(t))) \tag{4.10}$$

which only differs from (4.1) by a factor 2. This factor difference motivates us to introduce a variant of the dual extrapolation:

$$(x, y, s_+) \iff \begin{cases} x = Prox_{\overline{x}}^{\phi}(s)\\ y = Prox_{x}^{\phi}(\lambda g(x))\\ s_+ = s + \frac{\lambda}{2} g(y) \end{cases} \tag{4.11}$$

where

$$Prox_z^{\phi}(s) := arg\min_x \langle s, x\rangle + \omega^{\phi}(z, x).$$

To justify why (4.1) is a special case for (4.11), the following observation is crucial:

Lemma 4.12. $T_1^{\phi}(z, -s) = Prox_z^{\phi}(s)$.

Proof. It's shown in Lemma 4.4 that $T_1(z, -s) = \Phi(\nabla\phi(z) - s)$, we only have to verify that the right-hand side is the same thing.

$$\begin{aligned}
Prox_z^{\phi}(s) &= arg\min_x\{\langle s, x\rangle + \omega^{\phi}(z, x)\}\\
&= arg\min_x\{\langle s, x\rangle + \phi(x) - \phi(z) - \langle\nabla\phi(z), x - z\rangle\}\\
&= arg\min_x\{\langle s - \nabla\phi(z), x\rangle + \phi(x)\\
&= arg\max_x\{\langle\nabla\phi(z) - s, x\rangle - \phi(x)\\
&= \Phi(\nabla\phi(z) - s).
\end{aligned}$$

□

Therefore, if we pick s_t to be $-Jz_t$ and everything else stays the same, we will obtain the following recursive relationship:

$$-z(t+1) = -z(t) - \Phi(Jz(t) + 2J\Phi(Jz(t))) \tag{4.13}$$

which is exactly Algorithm 1.

4.2. **Convergence analysis for modified dual extrapolation.** We will try to accomplish two things in this section. Firstly, we will introduce relaxed relative lipschitzness as a generalization of area-convexity and see how relaxed relative lipschitzness gives us a guarantee of the convergence rate of (4.11). Secondly, we will see how that is linked to the conclusion of the Theorem 3.9.

Definition 4.14. We say that the an operator $J : \mathcal{C} \to \mathcal{C}^*$ is $\frac{1}{\eta}$-relaxed relatively lipschitz with respect to $\phi : \mathcal{C} \to \mathbb{R}$ if for all $a, b, c \in \mathcal{C}$, we have

$$\eta\langle J(b-a), b-c\rangle \leq \omega^\phi(a,b) + \omega^\phi(b,c) + \omega^\phi(a,c). \tag{4.15}$$

This is a generalization of both relative lipschitzness and area convexity. Relative lipschitzness is defined as follows:

Definition 4.16. We say that the an operator $J : \mathcal{C} \to \mathcal{C}^*$ is $\frac{1}{\eta}$ relatively lipschitz with respect to $\phi : \mathcal{C} \to \mathbb{R}$ if for all $a, b, c \in \mathcal{C}$, we have

$$\eta\langle J(b-a), b-c\rangle \leq \omega^\phi(a,b) + \omega^\phi(b,c). \tag{4.17}$$

Since Bregman's divergence is always non-negative, we can see that being $\frac{1}{\eta}$-relative Lipschitz implies being $\frac{1}{\eta}$-relaxed relative Lipschitz. The fact that Area-convexity implies relaxed relative lipschitzness follows from the following observation:

$$\begin{aligned}\phi(a) + \phi(b) + \phi(c) - 3\phi(\frac{a+b+c}{3}) &= \omega^\phi(a,b) + \omega^\phi(a,c) - 3\omega^\phi(a, \frac{a+b+c}{3}) \\ &\leq \omega^\phi(a,b) + \omega^\phi(a,c).\end{aligned}$$

After establishing this generalization of both relative lipschitzness and area convexity, we can see how using this generalization in place of relative lipschitzness grants a similar guarantee for the rate of convergence for our introduced variant of dual extrapolation (4.11) and therefore Algorithm 1.

Proposition 4.18. *(Rate of convergence) If g is $\frac{1}{\lambda}$-relaxed relatively Lipschitz with respect to ϕ, then we have the following guarantee on the convergence rate of (4.11): For all $u \in C$, we have that*

$$\sum_{0 \leq t < T} \langle g(y_t), y_t - u\rangle \leq \frac{2}{\lambda}\omega^\phi(\overline{x}, u). \tag{4.19}$$

To prove this, we start with a well-known lemma:

Lemma 4.20. *For $\omega = Prox_z^r(g), \forall u \in C, \langle g, w - u\rangle \leq \omega^r(z,u) - \omega^r(w,u) - \omega^r(z,w)$.*

Now we proceed with the proof of the proposition:

Proof. We first show that for each t,

$$\frac{\lambda}{2}\langle g(y_t), y_t - \overline{x}\rangle \leq \langle s_{t+1}, x_{t+1} - \overline{x}\rangle + \omega^\phi(\overline{x}, x_{t+1}) - \langle s_t, x_t - \overline{x}\rangle - \omega^\phi(\overline{x}, x_t). \tag{4.21}$$

To see this, we first apply the above lemma to the two proxy steps in the algorithm and get:

$$\langle s_t, x_t - x_{t+1}\rangle \leq \omega^\phi(\overline{x}, x_{t+1}) - \omega^\phi(x_t, x_{t+1}) - \omega^\phi(\overline{x}, x_t)$$
$$\frac{\lambda}{2}\langle g(x_t), y_t - x_{t+1}\rangle \leq \frac{\omega^\phi(x_t, x_{t+1}) - \omega^\phi(x_t, y_t) - \omega^\phi(y_t, x_{t+1})}{2}.$$

Additionally, by $\frac{1}{\lambda}-$relaxed relative Lipschitzness, we have:

$$\frac{\lambda}{2}\langle g(y_t) - g(x_t), y_t - x_{t+1}\rangle \leq \frac{\omega^\phi(y_t, x_{t+1}) + \omega^\phi(x_t, y_t) + \omega^\phi(x_t, x_{t+1})}{2}.$$

Now, if we add the three equations together, we get that

$$\langle s_t, x_t - x_{t+1}\rangle + \frac{\lambda}{2}\langle g(y_t), y_t - x_{t+1}\rangle \leq \omega^\phi(\overline{x}, x_{t+1}) - \omega^\phi(\overline{x}, x_t).$$

With our algorithm updating s: $s_{t+1} = s_t + \frac{\lambda}{2}g(y_t)$, we can see that:

$$\frac{\lambda}{2}\langle g(y_t), y_t - \overline{x}\rangle \leq \langle s_{t+1}, x_{t+1} - \overline{x}\rangle + \omega^\phi(\overline{x}, x_{t+1}) - \langle s_t, x_t - \overline{x}\rangle - \omega^\phi(\overline{x}, x_t)$$

which is what we want to see.

Now, an immediate result of this is that

$$A_t = \frac{\lambda}{2}\sum_{k=0}^{t-1}\langle g(y_k), y_k - \overline{x}\rangle - \langle s_t, x_t - \overline{x}\rangle - \omega^\phi(\overline{x}, x_t) \tag{4.22}$$

is non-increasing in t by considering the difference between A_t and A_{t-1}.

We are now ready to prove the proposition: For any $u \in C$

$$\begin{aligned}
\sum_{0\leq t<T}\langle g(y_t), y_t - u\rangle &= \sum_{0\leq t<T}\langle g(y_t), y_t - \overline{x}\rangle + \sum_{0\leq t<T}\langle g(y_t), \overline{x} - u\rangle + (\frac{2}{\lambda}\omega^\phi(\overline{x}, u) - \frac{2}{\lambda}\omega^\phi(\overline{x}, u)) \\
&\leq \sum_{0\leq t<T}\langle g(y_t), y_t - \overline{x}\rangle + \sum_{0\leq t<T}\langle g(y_t), \overline{x} - x_T\rangle + (\frac{2}{\lambda}\omega^\phi(\overline{x}, u) - \frac{2}{\lambda}\omega^\phi(\overline{x}, x_T)) \\
&\quad \text{(by the definition of } x_T) \\
&= \frac{2}{\lambda}A_T + \frac{2}{\lambda}\omega^\phi(\overline{x}, u) \\
&\leq \frac{2}{\lambda}A_0 + \frac{2}{\lambda}\omega^\phi(\overline{x}, u) \\
&= \omega^\phi(\overline{x}, u) \text{ (Notice that } x_0 = Prox_{\overline{x}}^\phi(0) = \overline{x}).
\end{aligned}$$

□

Remark 4.23. From this, we can see that moving from relative Lipschitzness to relaxed relative Lipschitzness poses a factor 2 difference to the dual extrapolation algorithm (which corresponds to taking 2 steps in the inner part of the update in Algorithm 1), and also a factor 2 to the rate of convergence.

We now move to try to understand this result in relation to the result in Sherman's paper. We first need to try to understand what is the $\sum_{0\leq t<T} g(y_t)$. This can be made clear by a very straightforward observation:

$$s_T = s_0 + \frac{\lambda}{2}\sum_{0\leq t<T} g(y_t) = \frac{\lambda}{2}\sum_{0\leq t<T} g(y_t) \tag{4.24}$$

when we set $s_0 = 0$. Therefore, when we let $g = J$ to be self-dual, we can restate Proposition 4.18 as:

$$\langle s_T, -u\rangle \leq \omega^\phi(\overline{x}, u) = \phi(u) - \phi(\overline{x}) - \langle \nabla\phi(\overline{x}), u - \overline{x}\rangle, \tag{4.25}$$

which serves as a bridge between the conclusion in Sherman's paper and this general claim of the rate of convergence of dual extrapolation.

We can reframe the conclusion of Theorem 3.9 to be (if Φ is a $0-$AMO):

$$\phi^*(-s_T) \leq \phi^*(0), \tag{4.26}$$

i.e., for all $u \in C$,

$$\langle s_T, -u\rangle \leq \phi(u) - \phi^*(0). \tag{4.27}$$

Notice if we assume $\nabla\phi(\overline{x}) = 0$, then the conclusion in Sherman is stronger than that for Proposition 4.18.

Using this analogy, we hope that we can modify dual extrapolation in such a way that uses not the oracle of $0-$AMO but a $\delta-$AMO. Before this, we establish the following lemma:

Lemma 4.28. *Given $k := \Phi(\nabla\phi(z) - g)$ where Φ is a $\delta -$ AMO for a convex ϕ, $w = Prox_z^\phi(g)$ and assume that for all z, g, $\forall u \in C$, we have that $\langle \nabla\phi(w) - \nabla\phi(k), u - k\rangle \leq 0$, we then have that $\forall u \in C$*

$$\langle g, k - u\rangle \leq \omega^\phi(z, u) - \omega^\phi(k, u) - \omega^\phi(z, k) + \delta. \tag{4.29}$$

Proof. By definition of k and $\delta-$AMO, we have that

$$\langle \nabla\phi(z) - g, k\rangle - \phi(k) + \phi(w) + \delta \geq \langle \nabla\phi(z) - g, w\rangle. \tag{4.30}$$

Similar to the proof of Lemma 4.20, we begin by considering the first-order optimality condition for $\langle g, w\rangle + \omega^\phi(z, w)$: $\forall u \in C$

$$\langle g + \nabla(\omega^\phi(z, w)), u - w\rangle \geq 0$$
$$\langle g + \nabla\phi(w) - \nabla\phi(z), u - w\rangle \geq 0$$
$$\langle \nabla\phi(z) - g, w - u\rangle + \langle \nabla\phi(w), u - w\rangle \geq 0$$
$$\langle \nabla\phi(z) - g, k - u\rangle - \phi(k) + \phi(w) + \delta + \langle \nabla\phi(w), u - w\rangle \geq 0 \text{ (by (3.26))}$$
$$\langle \nabla\phi(z) - g, k - u\rangle - (\phi(w) + \langle \nabla\phi(w), k - w\rangle) + \phi(w) + \delta + \langle \nabla\phi(w), u - w\rangle \geq 0 \text{ (by convexity of } \phi)$$
$$\langle \nabla\phi(z) - g, k - u\rangle + \langle \nabla\phi(w), u - k\rangle + \delta \geq 0$$
$$\langle g + \nabla(\omega^\phi(z, k)), u - k\rangle + \langle \nabla\phi(w) - \nabla\phi(k), u - k\rangle + \delta \geq 0$$
$$\langle g + \nabla(\omega^\phi(z, k)), u - k\rangle + \delta \geq 0 \text{ (by the assumption of the lemma).}$$

Therefore, we have that $\langle g, k - u\rangle \leq \langle -\nabla(\omega^\phi(z, k)), k - u\rangle + \delta$ which implies the lemma by Bregman distance cosine rule. □

With this lemma, we can control A_t up to an uncertainty of $O(\delta)$, precisely

Lemma 4.31. *If we have that for all z, g, $\forall u \in C$, it holds that $\langle \nabla\phi(w) - \nabla\phi(k), u - k\rangle \leq 0$, ϕ is convex, and g is $\frac{1}{\lambda}$ relatively Lipschitz with respect to ϕ, then updating according to (4.11) but replaced with Φ being a $\delta-$AMO gives us*

$$A_t \leq A_{t-1} + \frac{3}{2}\delta. \tag{4.32}$$

Proof. Apply an identical proof that's presented for Proposition 4.18, we have that

$$\langle s_t, x_t - x_{t+1}\rangle + \frac{\lambda}{2}\langle g(y_t), y_t - x_{t+1}\rangle \leq \omega^\phi(\overline{x}, x_{t+1}) - \omega^\phi(\overline{x}, x_t) + \frac{3}{2}\delta$$

which yields what we want. □

A direct result of this, without surprise, is that this updated version of dual extrapolation achieves an $\epsilon + \frac{3}{2}\delta$ approximate solution in $O(\epsilon^{-1})$ time.

Proposition 4.33. *With the same assumption as Lemma 4.28, then*

$$\sum_{0\leq t<T}\langle g(y_t), y_t - u\rangle \leq \frac{2}{\lambda}\omega^\phi(\overline{x}, u) + \frac{3T}{2}\delta.$$

4.3. **Energy conservation perspective.** Similarly to the Hamiltonian perspective of Algorithm 1, we want to close this section by identifying a candidate for the "Hamiltonian" in dual extrapolation. The idea is that it suffices to make this Hamiltonian conserved to guarantee the rate of convergence.

A natural candidate for such would be A_t from (4.22) since in the proof, we can easily see that if A_t is controlled to be not increased when updating, we have the guarantee of the rate of convergence. We will see how this is linked to the Hamiltonian $H(z) = \phi^*(Jz) = \phi^\omega(z)$ proposed in Section 3.

Let's take a look at A_t, when we set $g = J$ and $\nabla\phi(\overline{x}) = 0$, we have that

$$\begin{aligned} A_t &= \frac{\lambda}{2}\sum_{k=0}^{t-1}\langle g(y_k), y_k - \overline{x}\rangle - \langle s_t, x_t - \overline{x}\rangle - \omega^\phi(\overline{x}, x_t) \\ &= \langle s_t, -\overline{x}\rangle - \langle s_t, x_t - \overline{x}\rangle - \phi(x_t) + \phi(\overline{x}) + \langle \nabla\phi(\overline{x}), x_t - \overline{x}\rangle \\ &= \langle s_t, -x_t\rangle - \phi(x_t) + \phi(\overline{x}). \end{aligned}$$

On the other hand,

$$\begin{aligned} H(s_t) &= \phi^*(-s_t) \\ &= \langle -s_t, \Phi(-s_t)\rangle - \phi(\Phi(-s_t)) \\ &= \langle s_t, -x_t\rangle - \phi(x_t). \end{aligned}$$

This is essentially the same as A_t, which serves as a good indication that we might be able to view A_t or some variant of A_t as the Hamiltonian for dual extrapolation. If g is set to be a self-dual J, then it makes sense to make the Hamiltonian to be exactly the symplectic conjugate of ϕ in terms of J as that in Section 3. However, if g is more generalized, we would have to deal with the term $\frac{\lambda}{2}\sum_{k=0}^{t-1}\langle g(y_k), y_k\rangle$, and this is not easy to deal with, and it's even hard to assign it with a continuous correspondence. Therefore, we can now only claim that when we are dealing with g that satisfies $\langle g(x), x\rangle = 0$ (for example, under the self-dual case), we can let H defined as the symplectic conjugate be the hamiltonian and the dual extrapolation is just a discretization of the corresponding Hamiltonian system. Moreover, the conservation of Hamiltonian directly leads to the guarantee of the accelerated convergence rate.

Acknowledgments

I'd like to express my gratitude to my mentors, Antares Chen and Professor Lorenzo Orecchia, for their invaluable guidance throughout the topic selection, research process, and paper writing.

I'm also thankful to Professor Peter May for organizing the exceptional Math REU program. This program has allowed me to achieve more than I ever thought possible within a short timeframe.

Lastly, I want to extend my heartfelt appreciation to my family and friends for their unwavering support, without which I wouldn't have been able to complete this program or anything else.

References

[1] Beck, A. (2017d). First-order methods in optimization. Society for Industrial and Applied Mathematics; Mathematical Optimization Society.

[2] Bergmann, R., Herzog, R., Silva Louzeiro, M., Tenbrinck, D., & Vidal-Núñez, J. (2021). Fenchel Duality theory and a primal-dual algorithm on Riemannian manifolds. Foundations of Computational Mathematics, 21(6), 1465–1504. https://doi.org/10.1007/s10208-020-09486-5

[3] Cohen, Michael B., Sidford, Aaron, & Tian, Kevin. (2021). Relative Lipschitzness in extragradient methods and a direct recipe for acceleration. LIPIcs - Leibniz International Proceedings in Informatics. 12th Innovations in Theoretical Computer Science Conference (ITCS 2021).

[4] Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations. (2003). Computers & Mathematics with Applications, 45(10–11), 1782–1784. https://doi.org/10.1016/s0898-1221(03)80155-4

[5] Hamill, P. (2018). A student's guide to Lagrangians and Hamiltonians. Cambridge University Press.

[6] Inhoudsopgave - science.uu.nl Project CSG. (n.d.-a). https://webspace.science.uu.nl/~kolk0101/homepageHD/m.pdf

[7] Jambulapati, A., & Tian, K. (2023). Revisiting Area Convexity: Faster Box-Simplex Games and Spectrahedral Generalizations. https://doi.org/10.48550/arXiv.2303.15627

[8] Nemirovski, A. (2004). Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM Journal on Optimization, 15(1), 229–251. https://doi.org/10.1137/s1052623403425629

[9] Nesterov, Y. (2003a). Dual extrapolation and its applications for solving variational inequalities and related problems'. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.988671

[10] Phy411 lecture notes – canonical transformations. (n.d.-a). https://astro.pas.rochester.edu/ aquillen/phy411/lecture2.pdf

[11] Sherman, J. (2017). Area-convexity, l_∞ regularization, and undirected multicommodity flow. Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. https://doi.org/10.1145/3055399.3055501

[12] Taylor, M. E. (2011). Partial differential equations I. Applied Mathematical Sciences. https://doi.org/10.1007/978-1-4419-7055-8

[13] Teel, A. R., Poveda, J. I., & Le, J. (2019a). First-order optimization algorithms with resets and hamiltonian flows. 2019 IEEE 58th Conference on Decision and Control (CDC). https://doi.org/10.1109/cdc40024.2019.9029333