ACCELERATION FOR AFFINE-COUPLING PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Saddle-point problems play an important role in modern machine learning, including robust optimization and algorithmic fairness. We investigate structured convex–concave minimax problems where the primal variable lies in a high-dimensional, unconstrained space $\mathbb{R}^d$, while the dual variable is confined to a constrained, low-dimensional space $\mathbb{R}^J$, with $J \ll d$. Although existing theoretical frameworks establish that an accelerated rate of $\mathcal{O}(L/T^2)$ is possible, solving the associated subproblems efficiently can become a computational bottleneck in high dimensions. To address this, we introduce a decoupling technique. By restricting second-order updates to the low-dimensional dual space and employing first-order methods in the high-dimensional primal space, our framework combines Newton methods with Nesterov acceleration. For simplex-constrained dual variables and regularizers compatible with a self-concordant barrier, the resulting algorithm achieves a convergence rate of $\mathcal{O}(L/T^2)$ in primal objective suboptimality after T outer iterations, without strong dual concavity. The additional linear-algebra cost per outer iteration is $\widetilde{\mathcal{O}}(dJ^2 + J^{3.5})$, alongside one evaluation of the component losses and their Jacobian.

1 INTRODUCTION

We study minimax optimization problems of the form

$$\min_{w\in\mathbb{R}^d} \max_{\lambda\in\Lambda} \Phi(\lambda, w), \qquad \Phi(\lambda, w) := \langle \lambda, f(w)\rangle - R(\lambda). \tag{1}$$

Here $f : \mathbb{R}^d \to \mathbb{R}^J$ is smooth, $\Lambda \subset \mathbb{R}^J$ is compact and convex, and $R : \Lambda \to \mathbb{R}$ is convex. We call this structure *affine coupling*: the coupling term is linear in the typically low-dimensional variable λ, whereas f may depend nonlinearly on the high-dimensional variable w. Under Assumption 1.1, Φ is convex in w and concave in λ.

Problems of this form arise naturally when several component losses must be aggregated. Let $f_j(w)$ denote the loss associated with component j. Choosing $\Lambda = \Delta_J$ and $R(\lambda) = \frac{\gamma}{2}\|\lambda - \frac{1}{J}\mathbf{1}\|_2^2$ gives

$$\min_{w\in\mathbb{R}^d} \max_{\lambda\in\Delta_J} \left\{ \sum_{j=1}^{J} \lambda_j f_j(w) - \frac{\gamma}{2} \left\| \lambda - \frac{1}{J}\mathbf{1} \right\|_2^2 \right\}.$$

For $\gamma = 0$, the inner maximum is $\max_j f_j(w)$; increasing γ penalizes concentrated weights and moves the objective toward the uniform average $J^{-1}\sum_j f_j(w)$. When j indexes demographic groups or classes, this yields objectives used in fair and class-imbalanced learning, for example with $f_j(w) = \mathbb{E}[\ell(X, Y; w) \mid Y = j]$ (Hashimoto et al., 2018; Martinez et al., 2020). When j indexes environments or perturbation models, it yields a robust learning objective (Sagawa et al., 2020; Madry et al., 2018). Divergence-based and entropic regularizers yield related objectives in distributionally robust and entropy-regularized learning (Namkoong & Duchi, 2017; Hu & Hong, 2013; Ziebart et al., 2008; Haarnoja et al., 2018; Christodoulou, 2019; Geist et al., 2019).

From an optimization perspective, affine coupling lies between two familiar regimes. Generic first-order methods for convex–concave saddle-point problems have a worst-case convergence rate of order $1/T$ (Nesterov, 2018; Beck & Teboulle, 2012). When the dual variable disappears, for example, when $\Lambda = \Delta_1 = \{1\}$, the problem reduces to smooth convex minimization, for which acceleration yields the optimal rate $1/T^2$ (Nesterov, 1983). This raises our central question: can affine coupling

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

recover the accelerated rate without strong concavity, while remaining computationally efficient when $J \ll d$?

To answer this question, we design an algorithm that achieves the optimal accelerated convergence rate of $\mathcal{O}(1/T^2)$ while maintaining a per-iteration complexity that scales gracefully to high-dimensional settings. Specifically, our contributions are threefold:

1. Decoupled acceleration framework. Assuming oracle access to f and its Jacobian, we propose an algorithm achieving an $O(1/T^2)$ convergence rate for affine-coupling saddle-point problems (1). By leveraging a fixed-point structure to decouple the high-dimensional primal acceleration w from the low-dimensional dual constraints λ, we restrict exact second-order updates exclusively to the dual space. This bypasses the traditional bottleneck in high dimensions, reducing the per-iteration linear-algebra complexity from $\mathcal{O}((d+J)^3)$ for a joint dense solve to strictly $\mathcal{O}(dJ^2 + J^3)$.

2. Connecting interior-point methods and acceleration. We relate the local Newton decrement used in classical interior-point methods (Nesterov, 2018; Boyd & Vandenberghe, 2004) to the global approximation error required by the accelerated outer method (Nesterov, 1983). This theoretical bridge formally establishes that the nested inner problems can be solved only approximately while perfectly retaining the $O(1/T^2)$ convergence rate, translating theoretical composite acceleration into a computationally tractable algorithm.

3. Extending interior-point guarantees for general geometries. To rigorously bound the approximation error required by our first-order method, we extend the theoretical analysis of the classical path-following algorithm introduced by Nesterov & Nemirovski (1994). Specifically, we derive a novel analytical bound on the global cross-evaluation gap that holds for any compact convex constraint set Λ equipped with a self-concordant barrier.

1.1 Literature review

For general smooth convex-concave problems, optimistic and extragradient methods (Rakhlin & Sridharan, 2013; Mokhtari et al., 2020) and smoothing-based methods (Beck & Teboulle, 2012) achieve $\mathcal{O}(1/T)$ rates. These match worst-case first-order lower bounds when the dimensions are sufficiently large relative to the number of oracle calls (Ouyang & Xu, 2018). Faster rates are available under strong convexity or concavity (Thekumprampil et al., 2019), and for the two-objective case $R = 0$, $J = 2$ (Lezane, 2025).

Breaking the $\Omega(1/T)$ lower bound in bilinear games requires the dual update to utilize non-first-order information (Ouyang & Xu, 2018). Composite optimization provide a potential pathway to acceleration for our problem class. Indeed, (1) can be written as

$$\min_w h(f(w)), \qquad h(z) = \max_{\lambda \in \Lambda} \{\langle \lambda, z \rangle - R(\lambda)\}.$$

Following this perspective, Drusvyatskiy & Paquette (2017) propose an accelerated prox-linear method that theoretically achieves an $\mathcal{O}(1/T^2)$ outer convergence rate under standard convexity conditions. However, their algorithm requires repeatedly solving d-dimensional regularized primal subproblems to high precision. When f is nonlinear, finding these high-precision solutions is highly non-trivial. Employing general higher-order methods to solve these subproblems incurs prohibitive per-iteration costs, rendering the approach computationally intractable for large dimensions d. Table 1 summarizes the theoretical convergence rates of existing algorithms relative to our approach, while Section 5 provides a detailed accounting of both oracle and computational complexities.

1.2 Preliminaries

In this section, we introduce the preliminaries needed for our analysis.

Definition 1.1 (Bregman divergence)**.** *For a continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, the* Bregman divergence *from $x \in \mathbb{R}^d$ to $y \in \mathbb{R}^d$ is defined by*

$$D_F(x, y) := F(x) - F(y) - \langle \nabla F(y), x - y \rangle.$$

Definition 1.2 (L-Smoothness and γ-Strongly-convexity)**.** *A continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be L-*smooth*, if for all $x, y \in \mathbb{R}^d$*

$$D_F(x, y) \leq \frac{L}{2} \|x - y\|_2^2.$$

Setting	Existing method	Our method
Affine-coupling	$\mathcal{O}(\frac{L}{T})$ Rakhlin & Sridharan (2013)	
	$\mathcal{O}(\frac{L}{T})$ Mokhtari et al. (2020)	
	$\mathcal{O}(\frac{L}{T^2})^\dagger$ Drusvyatskiy & Paquette (2017)	$\mathcal{O}\left(\frac{L}{T^2}\right)$
R γ-Strongly-Convex	$\Omega(\frac{L}{\gamma T^2})$ Ouyang & Xu (2018)	
$R = 0, J = 2$	$\mathcal{O}(\frac{L\log(T)}{T^2})$ Lezane (2025)	

Table 1: Convergence rate comparison of various algorithms given T queries to ∇f. † whenever its subproblem oracle admits an efficient implementation.

F is said to be γ-strongly-convex*, if for all* $x, y \in \mathbb{R}^d$

$$D_F(x, y) \geq \frac{\gamma}{2}\|x - y\|_2^2.$$

Definition 1.3 (1-Self-concordance)**.** *A three times continuously differentiable convex function $F : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D} \subseteq \mathbb{R}^d$ is an open convex set, is said to be* 1-self-concordant *if for all $x \in \mathcal{D}$ and all directions $h \in \mathbb{R}^d$,*

$$|\nabla^3 F(x)[h, h, h]| \leq 2\left(h^\intercal \nabla^2 F(x) h\right)^{3/2},$$

where $\nabla^3 F(x)[h, h, h] := \frac{d^3}{dt^3} F(x + th)\Big|_{t=0}$ denotes the third-order directional derivative of F at x along the direction h.

Definition 1.4 (ν-Self-concordant barrier)**.** *A function $F : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D} \subseteq \mathbb{R}^d$ is an open convex set, is said to be a ν-*self-concordant barrier *for its closure $cl(\mathcal{D})$ if it is a 1-self-concordant function, $F(x) \to \infty$ as $x \to \partial\mathcal{D}$, and there exists a parameter $\nu \geq 0$ such that for all $x \in \mathcal{D}$ and all directions $h \in \mathbb{R}^d$,*

$$|h^\intercal \nabla F(x)| \leq \sqrt{\nu}\left(h^\intercal \nabla^2 F(x) h\right)^{1/2}.$$

Equivalently, when the Hessian $\nabla^2 F(x)$ is strictly positive definite, this condition can be written in terms of the Newton decrement as $\nabla F(x)^\intercal [\nabla^2 F(x)]^{-1} \nabla F(x) \leq \nu$.

Throughout this work, the affine-coupling problem is analyzed under the following assumptions.

Assumption 1.1 (Affine-Coupling)**.** *Consider the setting of* (1)*, and let $L, G \geq 0$.*

- *The set $\Lambda \subset \mathbb{R}^J$ is non-empty, compact, and convex. The primal and dual dimensions satisfy $d, J \in \mathbb{N}_+$, and our main computational regime of interest is $d \gg J$.*
- *There exists a ν-self-concordant barrier $B_\Lambda : \mathrm{relint}(\Lambda) \to \mathbb{R}$.*
- *The function $R : \Lambda \to \mathbb{R}$ is convex and R compatible with B_Λ (for every $c \geq 0$, $cR + B_\Lambda$ is convex and 1-self-concordant.)*
- *For every $\lambda \in \Lambda$, the function $w \mapsto f(w)\lambda$ is convex and L-smooth, and its gradient is bounded by G.*

2 Acceleration with a Fixed-Point Oracle

In this section we propose an algorithm and prove an accelerated convergence result for it. The construction uses an efficient first-order iterative scheme for w and a (non-constructive) fixed-point oracle for λ. This section serves as a warm up with clean analysis. Our main result, in the next section, tackles the challenge of efficiently implementing an approximate fixed point algorithm.

The following fixed-point property serves as the primary motivation for our design:

Theorem 2.1. *Consider any function $\Phi : \Lambda \times \mathbb{R}^d \to \mathbb{R}$ that is concave and continuous in the first argument and C^1 in the second argument. Let $\Lambda \subset \mathbb{R}^J$ be a non-empty, convex and compact subset. Then for any continuous update rule*

$$g_t : \begin{cases} \Lambda \to \mathbb{R}^d \\ \lambda \to g_t(\lambda), \end{cases}$$

there exists $\lambda_\star \in \Lambda$ such that

$$\Phi(\lambda_\star, g_t(\lambda_\star)) = \max_{\lambda \in \Lambda} \Phi(\lambda, g_t(\lambda_\star)).$$

The proof is provided in Section C. This theorem provides a crucial structural guarantee: if we treat $g_t(\lambda)$ as a primal update step, there always exists a dual solution $\lambda_t^\star$ that is optimal with respect to the exact primal variables it induces. To make this concrete and integrate it into our framework, we instantiate g_t as our primal optimization step. Specifically, for a fixed $\lambda_t^\star$, the gradient descent update rule is defined as:

$$g_t(\lambda) = w_{t+1}(\lambda) : \begin{cases} \Lambda \to \mathbb{R}^d \\ \lambda \to w_t - \eta_t \nabla f(w_t)\lambda \end{cases} .$$

with $\eta_t \geq 0$ the stepsize at iteration t. By embedding this structural property directly into the standard Nesterov acceleration framework, we obtain the following update rules for the primal variable w::

$$\begin{cases} w_t^{md} &= \frac{A_{t-1}}{A_t} w_t^{ag} + \frac{\alpha_t}{A_t} w_t \\ w_{t+1} &= w_t - \alpha_t \nabla f(w_t^{md}) \lambda_t^\star \\ w_{t+1}^{ag} &= \frac{A_{t-1}}{A_t} w_t^{ag} + \frac{\alpha_t}{A_t} w_{t+1} \end{cases} \tag{2}$$

The step-size parameters α_t and accumulation weights A_t follow a slight variation of the classical optimal sequences established in the accelerated optimization literature (Nesterov, 2018). With the initial condition $A_0 = 0$, and assuming L is the Lipschitz constant of the gradients defined later in the text, the parameter scheduling is governed by:

$$\begin{cases} \forall t \geq 1, \quad \alpha_t = \frac{1 + \sqrt{1 + 4LA_{t-1}}}{4L} \\ \forall t \geq 1, \quad A_t = A_{t-1} + \alpha_t. \end{cases} \tag{3}$$

Regarding the dual variable λ, we assume access to an oracle capable of computing the exact fixed point $\lambda_t^\star$ guaranteed by Theorem 2.1 at each iteration t. This fixed point satisfies the following condition:

$$\lambda_t^{\star\intercal} f\left(w_t^{md} - \frac{\alpha_t^2}{A_t} \nabla f(w_t^{md}) \lambda_t^\star\right) - R(\lambda_t^\star) = \max_{\lambda \in \Lambda} \left[\lambda^\intercal f\left(w_t^{md} - \frac{\alpha_t^2}{A_t} \nabla f(w_t^{md}) \lambda_t^\star\right) - R(\lambda)\right], \tag{4}$$

where we identify the new aggregated iterate as $w_{t+1}^{ag} = w_t^{md} - \frac{\alpha_t^2}{A_t} \nabla f(w_t^{md}) \lambda_t^\star$. We present the main convergence theorem.

Theorem 2.2. *Suppose that Assumption 1.1 holds. Following the update rules at Equations* (2) *to* (4)*, after T iterations, for all $w_\star \in \mathbb{R}^d$*

$$\max_{\lambda \in \Lambda} \Phi(\lambda, w_{T+1}^{ag}) - \max_{\lambda \in \Lambda} \Phi(\lambda, w_\star) = \mathcal{O}\left(\frac{L \|w_\star - w_{\text{init}}\|_2^2}{T^2}\right)$$

The full proof is deferred to Section D.1. In general, computing the exact fixed point $\lambda_t^\star$ at each step is computationally intractable. Consequently, in practical implementations, we solve an inner optimization problem to approximate this fixed point. We formalize this in the following sections, explicitly accounting for the resulting approximation errors in the final convergence rate

3 Main results

In this section we efficiently implement the fixed-point oracle approximately. To do so, we employ the Majorization-Minimization (MM) principle to formulate a sequence of surrogate subproblems, which we then solve using an interior-point method.

3.1 CROSS-EVALUATION GAP

Following the standard Majorization-Minimization (MM) framework (Lange et al., 2000; Yuille & Rangarajan, 2003), the natural surrogate $\tilde{g}_t : \mathbb{R}^d \to \mathbb{R}$ for minimizing an L-smooth function $g : \mathbb{R}^d \to \mathbb{R}$ at iteration t is given by:

$$\tilde{g}_t(w) := g(w_t^{md}) + \langle \nabla g(w_t^{md}), w - w_t^{md} \rangle + \frac{L}{2}\|w - w_t^{md}\|_2^2,$$

where w_t^{md} denotes the current iterate. The crucial step in extending this approach to saddle-point problems is parameterizing the update rule as $\lambda \to w_{t+1}^{ag}(\lambda)$ and deriving a surrogate $\tilde{\mu f}(w_{t+1}(\lambda), w_t^{md})$ for any $\mu \in \Lambda$ for all $\mu \in \Lambda$. Ultimately, our algorithm aims to find a $\lambda \in \Lambda$ that minimizes the gap between $\max_\mu \Phi(\mu, w_{t+1}^{ag}(\lambda))$ and $\Phi(\lambda, w_{t+1}^{ag}(\lambda))$.

To precisely quantify this difference, we introduce the cross-evaluation gap ϵ_t defined as:

$$\begin{aligned} \forall \lambda \in \Lambda, \quad S_t(\lambda) &:= \frac{\alpha_t^2}{2A_t}\|\nabla f(w_t^{md})\lambda\|_2^2 - \lambda^\intercal f(w_t^{md}) \\ \forall \lambda, \mu \in \Lambda, \quad \epsilon_t(\lambda, \mu) &:= \langle \nabla S_t(\lambda), \lambda - \mu \rangle + R(\lambda) - R(\mu). \end{aligned} \tag{5}$$

The following theorem characterizes the behavior of the saddle-point surrogate within the affine-coupling problem and specifies the class of inner subproblems that must be solved at each iteration.

Theorem 3.1 (Cross-evaluation gap)**.** *Suppose that Assumption 1.1 holds. We follow the update rules at Equation* (2)*, the cross-evaluation gap is strictly bounded by:*

$$\forall \lambda, \mu \in \Lambda, \quad \Phi(\mu, w_{t+1}^{ag}(\lambda)) - \Phi(\lambda, w_{t+1}^{ag}(\lambda)) \leq \epsilon_t(\lambda, \mu) + \frac{L}{2}\|w_{t+1}^{ag}(\lambda) - w_t^{md}\|_2^2$$

and by taking the maximum over μ, for any arbitrary $\lambda_{t+1} \in \Lambda$:

$$\begin{aligned} &\max_{\mu \in \Lambda} \Phi(\mu, w_{t+1}^{ag}(\lambda_{t+1})) - \Phi(\lambda_{t+1}, w_{t+1}^{ag}(\lambda_{t+1})) \\ \leq &\max_{\mu \in \Lambda} \epsilon_t(\lambda_{t+1}, \mu) + \frac{L}{2}\|w_{t+1}^{ag}(\lambda_{t+1}) - w_t^{md}\|_2^2. \end{aligned}$$

The subsequent theorem demonstrates that minimizing this surrogate directly yields the desired accelerated convergence rate.

Theorem 3.2 (Accelerated algorithm)**.** *Suppose that Assumption 1.1 holds. We follow the update rules at Equations* (2) *and* (3) *with λ_{t+1} verifying the follow property with $\epsilon_t \geq 0$:*

$$\max_{\lambda \in \Lambda} \Phi(\lambda, w_{t+1}^{ag}) - \Phi(\lambda_{t+1}, w_{t+1}^{ag}) \leq \epsilon_t + \frac{L}{2}\|w_{t+1}^{ag} - w_t^{md}\|_2^2 \tag{6}$$

Suppose that Equation (6) *holds for every iteration t. Then, after T iterations, we know that for all $w_\star \in \mathbb{R}^d$*

$$\max_{\lambda \in \Lambda} \Phi(\lambda, w_{T+1}^{ag}) - \max_{\lambda \in \Lambda} \Phi(\lambda, w_\star) = \mathcal{O}\left(\frac{L\|w_\star - w_{\text{init}}\|_2^2}{T^2} + \frac{1}{T^2}\sum_{t=1}^{T} t^2 \epsilon_t\right)$$

We defer the detailed proof to Section D.2 and Section E. Furthermore, the previous theorem highlights that accelerating convergence strictly relies on solving the inner problems to a precision of $\epsilon_t = \mathcal{O}(\frac{1}{t^4})$.

4 INTERIOR-POINT METHOD AND CROSS-EVALUATION GAP

At each outer iteration t, we solve the following convex inner minimization subproblem:

$$\min_{\lambda \in \Lambda} h_t(\lambda) := S_t(\lambda) + R(\lambda).$$

where S_t is a quadratic function define in Equation (5). Since S_t and R are both assumed to be convex, the following inequality holds for all $\lambda, \mu \in \Lambda$:

$$(S_t + R)(\lambda) - (S_t + R)(\mu) \leq \langle \nabla S_t(\lambda), \lambda - \mu \rangle + R(\lambda) - R(\mu) \leq \langle \nabla (S_t + R)(\lambda), \lambda - \mu \rangle. \tag{7}$$

Taking the maximum over $\mu \in \Lambda$ throughout this inequality naturally connects three fundamental quantities of the subproblem:

- **Optimality gap**: $\max_{\mu\in\Lambda}\Big[(S_t+R)(\lambda)-(S_t+R)(\mu)\Big]=h_t(\lambda)-\min h_t$
- **Cross-evaluation gap**: $\max_{\mu\in\Lambda}\Big[\langle\nabla S_t(\lambda),\lambda-\mu\rangle+R(\lambda)-R(\mu)\Big]=\max_{\mu\in\Lambda}\epsilon_t(\lambda,\mu)$
- **Frank-Wolfe gap**: $\max_{\mu\in\Lambda}\langle\nabla(S_t+R)(\lambda),\lambda-\mu\rangle=\max_{\mu\in\Lambda}\langle\nabla h_t(\lambda),\lambda-\mu\rangle$

The chain of inequalities in Equation (7) demonstrates that the optimality gap is upper-bounded by the cross-evaluation gap, which is itself bounded by the Frank-Wolfe gap. When R is self-concordant, the problem of minimising h_t is well-studied, allowing us to leverage standard algorithms and established convergence guarantees for the Frank-Wolfe gap. In such cases, these results can be directly applied to our cross-evaluation gap to ensure general convergence (see Section 4.1 for an example when R is quadratic).

However, in the general setting, R may be *compatible* (i.e., $cR+B$ is self-concordant) rather than strictly self-concordant. This is the case, for example, when $R(\lambda):=\sum_{j=1}^{J}\lambda_j\log(\lambda_j)$ is the negative Shannon entropy and $B_\Lambda=-\sum_{j=1}^{J}\log(\lambda_j)$ is the negative Burg entropy (See Lemma A.2).

To the best of our knowledge, the Frank-Wolfe gap evaluated at the output of standard interior-point solvers is not known to be tightly bounded. For instance, the path-following algorithm of Nesterov & Nemirovski (1994), which optimizes the barrier-penalized objective $h_t+\frac{B_\Lambda}{\tau}$, returns a solution $\lambda\in\Lambda$ satisfying

$$h_t(\lambda)-\min h_t\le\frac{\nu}{\tau},$$

but provides no theoretical guarantee for the Frank-Wolfe gap itself. Therefore, to extract a valid upper bound for the cross-evaluation gap ϵ_t, we must delve into the underlying mechanics of the path-following framework (Nesterov & Nemirovski, 1994) and explicitly exploit the quadratic structure of S_t.

In the remainder of this section, we first provide a concrete example of adapting an existing solver for our subproblem when R is quadratic (Section 4.1). We then establish theoretical guarantees for the more challenging general case of compatible regularizers in Section 4.2.

4.1 Standard Geometries and Optimality Certificates

In this section, we demonstrate how to integrate highly optimized, off-the-shelf solvers to address specific affine-coupling settings, culminating in a practical, end-to-end algorithmic solution. Specifically, when R is quadratic and Λ is a simplex, the inner subproblem reduces to a standard quadratic program. We can therefore delegate it to solvers as SciPy's `trust-constr`. Such solvers typically measure convergence via the inexact Karush-Kuhn-Tucker (KKT) optimality conditions of the constrained problem. To bridge this practical solver output with our theoretical guarantees, the following lemma translates these KKT tolerances into a formal bound on the Frank-Wolfe gap, thereby establishing an exact stopping criterion for our algorithm.

Lemma 4.1. *Suppose an interior-point algorithm terminates at a strictly feasible output $\mu_k\in$ relint(Δ_J), returning a barrier parameter $1/\tau_k\le\epsilon_\tau$ and dual multipliers $\nu_k\in\mathbb{R}$ and $z_k\in\mathbb{R}^J_+$ that satisfy the following inexact Karush-Kuhn-Tucker (KKT) conditions*

- ***Approximate Stationarity*** $||\nabla h_t(\mu_k)-\nu_k 1-z_k||_\infty\le\epsilon_g$
- ***Perturbed Complementary Slackness*** $\mu_{k,j}z_{k,j}=1/\tau_k$ *for all* $j\in\{1,...,J\}$

where ϵ_g is the tolerance bar for the stationary condition. Then, the Frank-Wolfe gap at μ_k is strictly bounded by:

$$\max_{\mu\in\Delta_J}\langle\nabla h_t(\mu_k),\mu_k-\mu\rangle\le J\epsilon_\tau+2\epsilon_g.$$

We defer the proof of this lemma to Section F. To ensure the Frank-Wolfe duality gap remains below the target precision $\frac{L}{t^4}$, we set the tolerances to $\epsilon_\tau:=\frac{L}{2Jt^4}$ and $\epsilon_g:=\frac{L}{4t^4}$. Given these choices, Boyd & Vandenberghe (2004) establish that a standard quadratic solver satisfies the stopping criteria in $\mathcal{O}\left(\sqrt{J}\log\left(\frac{1}{\epsilon_\tau}\right)\right)$ iterations. Consequently, Algorithm 1 yields an end-to-end algorithm achieving an $\mathcal{O}(1/T^2)$ accelerated convergence rate for the affine-coupling problem with a quadratic regularizer R over the simplex $\Lambda=\Delta_J$.

Algorithm 1 Accelerated Gradient Method with Quadratic surrogate

Initialize: Choose $w_1 \in \mathbb{R}^d$. Set L and T.
Set $A_0 = 0, w_1^{md} = w_1 = w_1^{ag}$
for $1 \le t \le T$ **do**
  Compute $\alpha_t = \frac{1+\sqrt{1+4LA_{t-1}}}{4L}, A_t = A_{t-1} + \alpha_t$
  Compute $w_t^{md} = \frac{A_{t-1}}{A_t} w_t^{ag} + \frac{\alpha_t}{A_t} w_t$
  Compute the Jacobian $\nabla f(w_t^{md})$ and the vector $f(w_t^{md})$
  Compute the surrogate $h_t(\lambda) := \frac{\alpha_t^2}{2A_t} \|\nabla f(w_t^{md})\lambda\|_2^2 - \lambda^\intercal f(w_t^{md}) + R(\lambda)$
  Call QP solver to minimize h_t in the simplex with tolerance rates $\left(\frac{L}{2Jt^4}, \frac{L}{4t^4}\right)$
    Obtain $\lambda_t^\star \in \Delta_J$
  Update $w_{t+1} = w_t - \alpha_t \nabla f(w_t^{md})\lambda_t^\star$
  Update $w_{t+1}^{ag} = \frac{A_{t-1}}{A_t} w_t^{ag} + \frac{\alpha_t}{A_t} w_{t+1}$
end for
return w_{T+1}^{ag}

4.2 Custom Path-Following Method for General Barriers

In this section, we analyze the underlying mechanics of the interior point solver to derive a strict upper bound for the cross-evaluation gap, ϵ_t. To formally establish convergence guarantees for the interior point method, we introduce the following condition on the optimal solution:

Assumption 4.1. *Let $w_\star \in \mathbb{R}^d$ be the global minimum. We assume that $\forall 1 \le i, j \le J$,*

$$|f_i(w_\star) - f_j(w_\star)| \le S_\star.$$

Following standard practices in interior point methods (Nesterov & Nemirovski, 1994), we consider the barrier-penalized objective parameterized by $\tau \ge 1$:

$$F_\tau(\lambda) := \tau(S_t(\lambda) + R(\lambda)) + B_\Lambda(\lambda). \tag{8}$$

Under our working assumptions, F_τ is strictly convex and 1-self-concordant. Minimizing this objective naturally yields the following Newton system, which we subsequently use to bound the cross-evaluation gap ϵ_t.

Definition 4.1 (Newton System and Decrement)**.** *Consider F_τ twice-continuously differentiable and a strictly feasible point $\mu \in \mathrm{relint}(\Lambda)$. Let $\{A\lambda + b\}$ be the Affine Hull of Λ. The Newton step direction $\Delta\mu \in \mathbb{R}^J$ and its associated local equality multiplier $s \in \mathbb{R}$ are defined as the solution to the symmetric linear system:*

$$\begin{bmatrix} \nabla^2 F_\tau(\mu) & A \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta\mu \\ s \end{bmatrix} = \begin{bmatrix} -\nabla F_\tau(\mu) \\ 0 \end{bmatrix} \tag{9}$$

When a solution exists, we quantify the proximity to the exact central path with Newton decrement:

$$\delta(\mu, F_\tau) := \sqrt{\Delta\mu^\intercal \nabla^2 F_\tau(\mu) \Delta\mu}. \tag{10}$$

Theorem 4.2 (Inner problem Upper bound via KKT block)**.** *Under Assumption 1.1, we consider a fixed $\tau \ge 1$ and $\mu_k \in$ relint(Λ). Suppose that the KKT block defined on (μ_k, F_τ) according to Definition 4.1 has a feasible solution $(\Delta\mu_k, s_k)$ and there exists $\beta \in [0, 1)$,*

$$\delta(\mu_k, F_\tau) \le \beta^2,$$

then for all $\lambda \in \Lambda$

$$\epsilon_t(\mu_k, \lambda) \le \frac{\nu}{\tau} + \frac{\sqrt{\nu}}{\tau}\left(\frac{\beta}{1-\beta}\right) + \frac{\omega_\star(\beta)}{\tau} + \frac{\beta}{1-\beta}\frac{2G\alpha_t}{\sqrt{\tau A_t}} \tag{11}$$

where $\omega_\star(\beta) := -\beta - \ln(1-\beta)$.

The preceding theorem connects standard interior point theory with the cross-evaluation bound essential to our acceleration framework. In Section G, we detail Algorithm 2, an adaptation of the path-following algorithm from (Nesterov & Nemirovski, 1994) tailored specifically to the simplex domain. The following corollary establishes a strict upper bound on the total number of iterations required by our inner loop algorithm.

Corollary 4.3. *Assume $\Lambda = \Delta_J$ and B_Λ to be negative Burg entropy. Suppose $D = \|w_\star - w_1\|_2, L \geq 1$. Let the parameters be set according to Equations* (2) *and* (3) *with $\beta = 1/4$ and $\epsilon_t := L/t^4$. Under Assumptions 1.1 and 4.1, the total number of inner loop iterations N_{Inner} in Algorithm 2 satisfies*

$$N_{Inner} = \mathcal{O}\left(G^2 + \sqrt{J}\log\left(JGt\right) + \frac{G^2}{L} + 2G(\sqrt{D^2 + \log(t)} + S_\star\right)$$

The proof and further theoretical details are deferred to Section G.1.

5 IMPLEMENTATION

Before entering the simulation, we estimate the overall computational complexity of our algorithm. During the inner loop optimization over λ, the Jacobian of f evaluated at w_t^{md} remains fixed. Consequently, each inner loop requires exactly one evaluation of ∇f and at most N_{inner} evaluations of the penalty derivatives $(\nabla R, \nabla^2 R)$. Unlike standard first-order methods, our second-order inner loop incurs the overhead of computing and inverting a $J \times J$ Hessian, adding $\mathcal{O}(dJ^2 + J^3 N_{\text{inner}})$ to the per-iteration cost. Factoring in the total iterations required to reach an ϵ-approximate saddle point, we summarize the overall complexities in Table 2 for a rigorous comparison with existing baselines.

Algorithms	**Call ∇f**	**Call $(\nabla R, \nabla^2 R)$**	**Computational cost**
Rakhlin & Sridharan (2013)	$\mathcal{O}\left(\frac{L}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\epsilon}dJ\right)$
Ouyang & Xu (2018)	$\mathcal{O}\left(\sqrt{\frac{L}{\gamma\epsilon}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\gamma\epsilon}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\gamma\epsilon}}dJ\right)$
Our methods	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\epsilon}}\left(dJ^2 + J^{3.5}\right)\right)$

Table 2: Overall comparison of various algorithms for the saddle point problem to reach precision ϵ. The $\tilde{\mathcal{O}}$ notation hides additional multiplicative logarithmic factors.

Synthetic Logistic Regression To empirically validate our theoretical guarantees, we evaluate the proposed algorithm on a synthetic group-structured classification task. Specifically, we consider a regularized minimax logistic regression problem. Suppose we have J demographic groups, where each group $j \in \{1, \ldots, J\}$ contains a dataset of $N = 5,000$ samples $(X_{j,i}, y_{j,i}) \in \mathbb{R}^d \times \{-1, 1\}$. The component loss $f_j(w)$ is defined as the empirical logistic risk for group j:

$$f_j(w) = \frac{1}{N}\sum_{i=1}^{N}\log\left(1 + \exp\left(-y_{j,i}X_{j,i}^\intercal w\right)\right)$$

$$\min_{w\in\mathbb{R}^d}\max_{\lambda\in\Delta_J}\left\{\sum_{j=1}^{J}\lambda_j f_j(w) - \frac{\gamma}{2}\left\|\lambda - \frac{1}{J}\mathbf{1}\right\|_2^2\right\}.$$

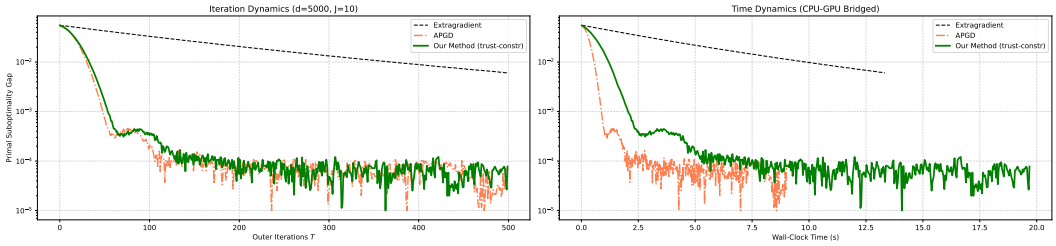
where Δ_J is the simplex and $\gamma \geq 0$ dictates the strong concavity of the regularizer. Our primary evaluation metrics are the iteration count and the GPU wall-clock time required to achieve an ϵ-suboptimal primal gap (full simulation details are provided in Section I). We benchmark our approach against established first-order methods—Accelerated Proximal Gradient Descent (APGD) and Extragradient (EG)—as well as higher-order solvers like L-BFGS-B. The evaluation spans three main axes, with overall results summarized in Section 5:

1. Robustness to the regularization parameter (γ). A key theoretical advantage of our method is its convergence guarantees even in the absence of strong concavity ($\gamma = 0$). Empirically, while APGD degrades significantly for small values of γ, our method and EG remain robust.

2. Accelerated high-precision convergence. Compared to standard first-order approaches, the primary benefit of our accelerated algorithm is its ability to reach high-precision solutions in a fraction of the time.

3. Scaling with dimensions (d and J). Although our higher-order implementation lacks low-level optimization, we compare its computational efficiency against SciPy's L-BFGS-B. Our algorithm exhibits superior scaling with respect to the primal dimension d, which offsets its theoretical complexity disadvantage in the dual dimension J.

(a) Convergence behavior of our algorithm (green) compared to APGD (orange) and EG (black). The left pane plots the primal gap against the iteration count, while the right pane plots it against computational wall-clock time. ($d = 5000, J = 10$)



(b) Time to reach precision $\epsilon = 10^{-2}$ ($d = 5000, J = 10$)

Algorithm	$\gamma = 0$	$\gamma = 0.001$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1$
Ours	34 (1.16s)	35 (1.46s)	34 (1.22s)	34 (1.24s)	34 (1.20s)
APGD	N/A	88 (1.59s)	32 (0.59s)	20 (0.36s)	18 (0.32s)
EG	210 (5.48s)	219 (5.92s)	372 (9.90s)	-/-	-/-

(c) Time to reach precision ϵ ($d = 5000, J = 30$)

Algorithm	$\epsilon = 10^{-1}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-3}$
Ours	1 (0.02s)	34 (1.42s)	55 (2.38s)
APGD	1 (0.02s)	32 (0.57s)	49 (0.88s)
EG	1 (0.03s)	372 (9.88s)	-/-

(d) Complexity scaling in d ($J = 10, \gamma = 10^{-2}, T = 200$)

	Ours		**L-BFGS-B**	
Dimension d	Obj.	Time (s)	Obj.	Time (s)
1000	0.675157	5.64	0.675161	0.13
2000	0.670539	6.74	0.670589	0.42
5000	0.637399	7.71	0.637422	1.07

(e) Complexity scaling in J ($d = 5000, \gamma = 10^{-2}, T = 200$)

	Ours		**L-BFGS-B**	
Dimension J	Obj.	Time (s)	Obj.	Time (s)
4	0.545715	5.28	0.545861	0.36
8	0.622306	6.50	0.622308	0.63
16	0.658493	9.81	0.658977	1.67
32	0.676453	16.84	0.676627	1.08

Table 3: Summary of experimental results across all settings.

AI USE STATEMENT

In this work, we used generative AI tools to help develop to provide critical ingredients for proving mathematical claims, to assist in the writing of proofs, generate synthetic data sets and to assist with translation. We have not used generative AI tools to theoretical models or conceptual framework, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods or interpret results. The categories "clean and reformat dataset" and "support qualitative and thematic data analysis" are not applicable to this work.

Additionally, we used generative AI tools to transcribe recordings of research material, summarize or analyse existing literature, edit a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our work, we provide an in-depth breakdown of the synthetic data generation and experimental configurations. On the theoretical side, rigorous and complete proofs supporting all claims, theorems, and bounds are detailed in the Appendix.

REFERENCES

Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012. doi: 10.1137/100818327. URL https://doi.org/10.1137/100818327.

C. Berge. *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity*. Dover books on mathematics. Dover Publications, 1997. ISBN 9780486696539. URL https://books.google.de/books?id=TyDeuKolqJIC.

S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Number Teil 1 in Berichte über verteilte messysteme. Cambridge University Press, 2004. ISBN 9780521833783. URL https://books.google.de/books?id=mYm0bLd3fcoC.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993. doi: 10.1137/0803026. URL https://doi.org/10.1137/0803026.

Petros Christodoulou. Soft actor-critic for discrete action settings. *arXiv preprint arXiv:1910.07207*, 2019. Explicitly adapts Maximum Entropy RL to discrete action spaces, formulating the policy as a categorical distribution over a probability simplex penalized by Shannon entropy.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps, 2017. URL https://arxiv.org/abs/1605.00125.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning (ICML)*, pp. 2160–2169. PMLR, 2019. Provides theoretical equivalence between Shannon entropy-regularized MDPs, convex duality, and softmax policies on the simplex.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pp. 1861–1870. PMLR, 2018. Introduces the Soft Actor-Critic algorithm, utilizing Shannon entropy to encourage exploration and prevent deterministic policy collapse.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938, 2018. URL https://proceedings.mlr.press/v80/hashimoto18a.html.

Zhaolin Hu and L Jeff Hong. Kl divergence distributionally robust optimization with application to data-driven inventory models. *Operations Research*, 61(4):981–996, 2013.

Shizuo Kakutani. A generalization of brouwer's fixed point theorem. *Duke Mathematical Journal*, 8: 457–459, 1941. URL https://api.semanticscholar.org/CorpusID:122550649.

Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000. doi: 10.1080/10618600.2000.10474858. URL https://doi.org/10.1080/10618600.2000.10474858.

Clément Lezane. *First Order Methods with Non-Euclidean Geometry*. Phd thesis - research ut, graduation ut, University of Twente, Netherlands, June 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6755–6764, 2020. URL https://proceedings.mlr.press/v119/martinez20a.html.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1497–1507. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/mokhtari20a.html.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 2975–2984, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983. URL https://api.semanticscholar.org/CorpusID:145918791.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.

Yurii Nesterov and Arkadii Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185:1 – 35, 2018. URL https://api.semanticscholar.org/CorpusID:52995339.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/f0dd4a99fba6075a9494772b58f95280-Paper.pdf.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

Kiran Koshy Thekumprampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. *Efficient algorithms for smooth minimax optimization*. Curran Associates Inc., Red Hook, NY, USA, 2019.

A. L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15 (4):915–936, 04 2003. ISSN 0899-7667. doi: 10.1162/08997660360581958. URL https://doi.org/10.1162/08997660360581958.

Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 8, pp. 1433–1438, 2008. The seminal work formally introducing the Maximum (Shannon) Entropy principle to resolve ambiguity in behavior distributions.

A Complementary materials

Lemma A.1. *Let f be a convex function and ν be convew and continuously differentiable. For*

$$u^\star = \arg\min_{u\in\mathcal{X}}\{f(u) + D^\nu(u, y)\},$$

and all u, it holds

$$f(u^\star) + D^\nu(u^\star, y) + D^f(u, u^\star) \leq f(u) + D^\nu(u, y) - D^\nu(u, u^\star).$$

Proof. As $u^\star$ is a minimizer, it fulfills for all $u \in \mathcal{X}$ the first-order optimality condition

$$\langle \nabla f(u^\star) + \nabla D^\nu(u^\star, y), u - u^\star\rangle \geq 0, \tag{12}$$

with the gradient taken with respect to the first entry. Applying the three-points identity Chen & Teboulle (1993), yields

$$\langle \nabla D^\nu(u^\star, y), u - u^\star\rangle = D^\nu(u, y) - D^\nu(u, u^\star) - D^\nu(u^\star, y),$$

leading to

$$f(u) - f(u^\star) - D^f(u, u^\star) = \langle \nabla f(u^\star), u - u^\star\rangle \geq D^\nu(u, u^\star) - D^\nu(u, y) + D^\nu(u^\star, y).$$

□

Lemma A.2 (Shannon Entropy)**.** *If R is the negative Shannon entropy and B is the negative Burg entropy, for all $c \geq 0$, $cR + B$ is 1-self-concordant.*

Proof. Let $c \geq 0$, the sum is:

$$g(\mu) = c\mu \ln \mu - \ln \mu$$

We calculate the derivatives

$$g''(\mu) = \frac{c}{\mu} + \frac{1}{\mu^2} = \frac{c\mu + 1}{\mu^2}$$
$$g'''(\mu) = -\frac{c}{\mu^2} - \frac{2}{\mu^3} = -\frac{c\mu + 2}{\mu^3}$$

Now, the 1-self-concordance ($|g'''(\mu)| \leq 2(g''(\mu))^{3/2}$) can be rewritten as :

$$\frac{c\mu + 2}{\mu^3} \leq 2\left(\frac{c\mu + 1}{\mu^2}\right)^{3/2} = \frac{2(c\mu + 1)^{3/2}}{\mu^3}$$

Since $\mu > 0$, we consider $y := c\mu \geq 0$ and multiply both sides by μ^3

$$c\mu + 2 \leq 2(c\mu + 1)^{3/2}$$
$$\iff \quad y + 2 \leq 2(y + 1)^{3/2}.$$

The inequality holds for $y = 0$. The left side grows at a constant rate of 1. The right side grows at a rate of $3(y + 1)^{1/2} \geq 3$. Because the right side grows faster than the left side, the inequality is strictly and globally true for all $y \geq 0$. □

B STEP-SIZES

We recall that the Nesterov accelerated gradient scheme:

$$\begin{cases} A_0 = 0 \\ \forall t \geq 1, \quad \alpha_t = \frac{1+\sqrt{1+4LA_{t-1}}}{4L} \\ \forall t \geq 1, \quad A_t = A_{t-1} + \alpha_t \end{cases}$$

Lemma B.1. *We have for all $t \geq 1$*

$$2L\alpha_t^2 \leq A_t$$
$$\frac{t^2+4t}{16L} < A_t \leq \frac{(t+2)^2}{16L} + \frac{(t+2)\ln(t+1)}{4L} + \frac{\ln^2(t+1)}{4L}$$

Proof. By the definition of α_t,

$$\begin{aligned} 4L\alpha_t - 1 &= \sqrt{1+4LA_{t-1}} \\ \implies (4L\alpha_t - 1)^2 &= 1 + 4LA_{t-1} \\ \implies 16L^2\alpha_t^2 - 8L\alpha_t + 1 &= 1 + 4LA_{t-1}. \end{aligned}$$

After rearranging the terms, we have:

$$\begin{aligned} 4L\alpha_t^2 - 2\alpha_t &= A_{t-1} = A_t - \alpha_t \\ \implies 4L\alpha_t^2 &= A_t + \alpha_t. \end{aligned}$$

For the first inequality, we obtain

$$\frac{L\alpha_t^2}{A_t} = \frac{A_t + \alpha_t}{4A_t} = \frac{1}{4}\left(1 + \frac{\alpha_t}{A_t}\right) \leq \frac{1}{4}(1+1) = \frac{2}{4} = \frac{1}{2}$$

which yields the first inequality. Now continue to develop

$$\begin{aligned} 4L\alpha_t^2 - 2\alpha_t &= A_{t-1} \\ \implies 4L^2\alpha_t^2 - 2L\alpha_t + \frac{1}{4} &= LA_{t-1} + \frac{1}{4} \\ \implies \left(2L\alpha_t - \frac{1}{2}\right)^2 &= LA_{t-1} + \frac{1}{4} \end{aligned}$$

Now, we take the square root and rearrange the terms

$$\begin{aligned} 2L\alpha_t - \frac{1}{2} &= \sqrt{LA_{t-1} + \frac{1}{4}} \\ \implies L\alpha_t &= \frac{1}{4} + \frac{1}{2}\sqrt{LA_{t-1} + \frac{1}{4}} \\ \implies LA_t + \frac{1}{4} &= \left(LA_{t-1} + \frac{1}{4}\right) + \frac{1}{2}\sqrt{LA_{t-1} + \frac{1}{4}} + \frac{1}{4} \end{aligned}$$

Let us introduce $x_t = \sqrt{LA_t + \frac{1}{4}}$ with $x_0 = \frac{1}{2}$, substituting this yields :

$$x_t^2 = x_{t-1}^2 + \frac{1}{2}x_{t-1} + \frac{1}{4}$$

For lower bound, we notice that for $t \geq 1$:

$$\begin{aligned} \left(x_{t-1} + \frac{1}{4}\right)^2 &= x_{t-1}^2 + \frac{1}{2}x_{t-1} + \frac{1}{16} \\ \implies x_t^2 &> \left(x_{t-1} + \frac{1}{4}\right)^2 \\ \implies x_t &> x_{t-1} + \frac{1}{4} \end{aligned}$$

By induction,

$$x_t > x_0 + \frac{t}{4} = \frac{t}{4} + \frac{1}{2}$$
$$x_t^2 > \frac{t^2}{16} + \frac{t}{4} + \frac{1}{4}$$
$$A_t > \frac{t^2 + 4t}{16L}$$

For upper bound Since x_t is increasing, we notice for $t \geq 1$

$$x_t - x_{t-1} = \frac{x_t^2 - x_{t-1}^2}{x_t + x_{t-1}} = \frac{\frac{1}{2}x_{t-1} + \frac{1}{4}}{x_t + x_{t-1}} < \frac{\frac{1}{2}x_{t-1} + \frac{1}{4}}{2x_{t-1}} = \frac{1}{4} + \frac{1}{8x_{t-1}}$$

We recall that $x_{t-1} \geq \frac{t-1}{4} + \frac{1}{2} = \frac{t+1}{4}$

$$x_t - x_{t-1} \leq \frac{1}{4} + \frac{1}{8\left(\frac{t+1}{4}\right)} = \frac{1}{4} + \frac{1}{2(t+1)}$$
$$\implies x_t - x_0 \leq \frac{t}{4} + \frac{1}{2}\sum_{k=1}^{t} \frac{1}{k+1}$$

Since $\left(\sum_{k=1}^{t} \frac{1}{k+1} < \int_1^{t+1} \frac{du}{u} = \ln(t+1)\right)$, we get:

$$x_t \leq \frac{t}{4} + \frac{1}{2} + \frac{1}{2}\ln(t+1)$$
$$x_t^2 \leq \frac{(t+2)^2}{16} + \frac{(t+2)\ln(t+1)}{4} + \frac{\ln^2(t+1)}{4}$$
$$A_t = \frac{x_t^2 - \frac{1}{4}}{L} \leq \frac{(t+2)^2}{16L} + \frac{(t+2)\ln(t+1)}{4L} + \frac{\ln^2(t+1)}{4L}.$$

□

C FIXED POINT THEOREM

Theorem C.1 (Berge's Maximum Theorem, Berge (1997))**.** *Let X and Y be topological spaces. Let $f : X \times Y \to \mathbb{R}$ be a continuous function, and let $C : Y \to 2^X$ be a continuous set-valued mapping (i.e. upper and lower hemicontinuous) with non-empty compact values. Then,*

- *The marginal maximum value function $M : Y \to \mathbb{R}$ defined by $M(y) = \max_{x \in C(y)} f(x, y)$ is continuous.*
- *The argmax correspondence $A : Y \to 2^X$ defined by $A(y) = \{x \in C(y) \mid f(x, y) = M(y)\}$ is upper hemicontinuous with non-empty and compact values.*

Theorem C.2 (Kakutani's Fixed-Point Theorem, Kakutani (1941))**.** *Let S be a non-empty, compact, and convex subset of a Euclidean space (or a locally convex topological vector space) $\mathbb{R}^n$. Let $\phi : S \to 2^S$ be a set-valued mapping. If ϕ is upper hemicontinuous (which implies a closed graph when S is compact) and $\phi(x)$ is non-empty and convex for all $x \in S$, then ϕ has a fixed point. That is, there exists an $x^\star \in S$ such that $x^\star \in \phi(x^\star)$.*

From the two previous Theorem, we can prove our Theorem.

Theorem C.3. *Suppose that $\Phi : \Lambda \times \mathbb{R}^d \to \mathbb{R}$ is concave in the first argument and $\mathcal{C}_1$ in the second argument. If $\Lambda \subset \mathbb{R}^J$ is a non empty, convex and compact subset, then for any continuous update rule*

$$g_t : \begin{cases} \Lambda \to W \\ \lambda \to g_t(\lambda) \end{cases}$$

there exists $\lambda_\star \in \Lambda$ such that

$$\Phi(\lambda_\star, g_t(\lambda_\star)) = \max_{\lambda \in \Lambda} \Phi(\lambda, g_t(\lambda_\star)).$$

Proof. We consider the best-response mapping

$$\forall \lambda \in \Lambda, \quad S(\lambda) := \arg\max_{\mu \in \Lambda} \Phi(\mu, g_t(\lambda))$$

Because the objective $\Phi(\lambda; w_t)$ and the continuous update rule $g_t(\lambda)$ are continuous, and the constraint set Λ is compact, Theorem C.1 guarantees that for all $\lambda \in \Lambda$, $S(\lambda)$ is non-empty, compact-valued, and upper hemicontinuous (possesses a closed graph).

We note that the mapping $S(\lambda)$ is convex because the function $\Phi(\cdot, w_t)$ is concave. Since $S(\lambda)$ is an upper hemicontinuous, convex-valued mapping from a compact, convex set to itself, Theorem C.2 guarantees it must have a fixed point $\lambda_\star \in S(\lambda_\star)$. □

D Proof for the Acceleration

D.1 Noiseless case

We recall the accelerated algorithm

$$\begin{cases} w_t^{md} & = \frac{A_{t-1}}{A_t} w_t^{ag} + \frac{\alpha_t}{A_t} w_t \\ & \lambda_t^\star \quad \text{Fixed point from Theorem 2.1} \\ w_{t+1} & = w_t - \alpha_t \nabla f(w_t^{md}) \lambda_t^\star \\ w_{t+1}^{ag} & = \frac{A_{t-1}}{A_t} w_t^{ag} + \frac{\alpha_t}{A_t} w_{t+1} \end{cases} \tag{13}$$

With the following fixed-point property

$$\lambda_t^{\star,\intercal} f(w_{t+1}^{ag}) - R(\lambda_t^\star) = \max_{\lambda \in \Lambda} [\lambda \cdot f(w_{t+1}^{ag}) - R(\lambda)]$$

Proof. At iteration t, we have $w_{t+1} = w_t - \alpha_t \nabla f(w_t^{md}) \lambda_t^\star$, therefore

$$\begin{aligned} \alpha_t \langle \nabla f(w_t^{md}) \lambda_t^\star, w_{t+1} - w_\star \rangle &= \langle w_t - w_{t+1}, w_{t+1} - w_\star \rangle \\ &= \frac{1}{2}\Big(\|w_\star - w_t\|_2^2 - \|w_\star - w_{t+1}\|_2^2 - \|w_{t+1} - w_t\|_2^2 \Big) \end{aligned}$$

where we use the identity $\langle a, b \rangle = \frac{1}{2}\left(\|a+b\|_2^2 - \|a\|_2^2 - \|b\|_2^2 \right)$. Then, we rearrange the term into

$$\begin{aligned} \alpha_t \langle \nabla f(w_t^{md}) \lambda_t^\star, w_t^{md} - w_\star \rangle = \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 \\ + \alpha_t \langle \nabla f(w_t^{md}) \lambda_t^\star, w_t^{md} - w_{t+1} \rangle - \frac{1}{2}\|w_{t+1} - w_t\|_2^2 \end{aligned}$$

We apply the convexity of $\lambda_t^\star \cdot f$ on w from Assumption 1.1

$$\begin{aligned} \lambda_t^{\star,\intercal} f(w_t^{md}) - \lambda_t^{\star,\intercal} f(w_\star) &\leq \langle \nabla f(w_t^{md}) \lambda_t^\star, w_t^{md} - w_\star \rangle \\ \alpha_t \Big(\lambda_t^{\star,\intercal} f(w_t^{md}) - \lambda_t^{\star,\intercal} f(w_\star) \Big) &\leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 - \frac{1}{2}\|w_{t+1} - w_t\|_2^2 \\ &\quad + \alpha_t \langle \nabla f(w_t^{md}) \lambda_t^\star, w_t^{md} - w_{t+1} \rangle \end{aligned} \tag{14}$$

After arranging, we have

$$\begin{aligned} \alpha_t \Big(\lambda_t^{\star,\intercal} f(w_t^{md}) + \langle \nabla f(w_t^{md}) \lambda_t^\star, w_{t+1} - w_t^{md} \rangle \Big) \leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 \\ + \alpha_t \lambda_t^{\star,\intercal} f(w_\star) - \frac{1}{2}\|w_{t+1} - w_t\|_2^2 \end{aligned} \tag{15}$$

On the left hand side, it is related to $\lambda_t^\star \cdot f(w_{t+1}^{ag})$. We use the L-smoothness of $\lambda_t^\star \cdot f$,

$$\lambda_t^\star \cdot f(w_{t+1}^{ag}) \leq \lambda_t^\star \cdot f(w_t^{md}) + \langle \nabla f(w_t^{md}) \lambda_t^\star, w_{t+1}^{ag} - w_t^{md} \rangle + \frac{L}{2}\|w_{t+1}^{ag} - w_t^{md}\|^2$$

From the fixed-point property $\lambda_t^{\star,\intercal} f(w_{t+1}^{ag}) + R(\lambda_t^\star) = \max_{\lambda\in\Lambda}[\lambda \cdot f(w_{t+1}^{ag}) + R(\lambda)]$, we know

$$\begin{aligned}\max_{\lambda\in\Lambda} \Phi(\lambda, w_{t+1}^{ag}) &= \lambda_t^{\star,\intercal} f(w_{t+1}^{ag}) - R(\lambda_t^\star) \\ &\leq \lambda_t^\star \cdot f(w_t^{md}) + \langle \nabla f(w_t^{md})\lambda_t^\star, w_{t+1}^{ag} - w_t^{md}\rangle \\ &+ \frac{L}{2}\|w_{t+1}^{ag} - w_t^{md}\|_2^2 - R(\lambda_t^\star)\end{aligned} \tag{16}$$

Now we make the following decomposition for the scalar product :

$$\langle \nabla f(w_t^{md})\lambda_t^\star, w_{t+1}^{ag} - w_t^{md}\rangle = \frac{\alpha_t}{A_t}\langle \nabla f(w_t^{md})\lambda_t^\star, w_{t+1} - w_t^{md}\rangle + \frac{A_{t-1}}{A_t}\langle \nabla f(w_t^{md})\lambda_t^\star, w_t^{ag} - w_t^{md}\rangle$$

Therefore, Equation (16) scaled by A_t becomes

$$\begin{aligned}A_t \max_{\lambda\in\Lambda} \Phi(\lambda, w_{t+1}^{ag}) &\leq \alpha_t \left(\lambda_t^{\star,\intercal} f(w_t^{md}) + \langle \nabla f(w_t^{md})\lambda_t^\star, w_{t+1} - w_t^{md}\rangle\right) \\ &+ A_{t-1}\left(\lambda_t^{\star,\intercal} f(w_t^{md}) + \langle \nabla f(w_t^{md})\lambda_t^\star, w_t^{ag} - w_t^{md}\rangle\right) \\ &+ \frac{LA_t}{2}\|w_{t+1}^{ag} - w_t^{md}\|_2^2 - A_t R(\lambda_t^\star)\end{aligned} \tag{17}$$

We recall Equation (15) which contains the first line in Equation (17)

$$\begin{aligned}\alpha_t\Big(\lambda_t^{\star,\intercal} f(w_t^{md}) + \langle \nabla f(w_t^{md})\lambda_t^\star, w_{t+1} - w_t^{md}\rangle\Big) &\leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 \\ &+ \alpha_t \lambda_t^\star \cdot f(w_\star) - \frac{1}{2}\|w_{t+1} - w_t\|_2^2\end{aligned}$$

The second term in Equation (17) is related to $\lambda_t^{\star,\intercal} f(w_t^{ag})$. By the convexity of $\lambda_{t+1} \cdot f$

$$A_{t-1}\left(\lambda_t^{\star,\intercal} f(w_t^{md}) + \langle \nabla f(w_t^{md})\lambda_t, w_t^{ag} - w_t^{md}\rangle\right) \leq A_{t-1}\lambda_t^{\star,\intercal} f(w_t^{ag})$$

Using both simplifications, Equation (17) becomes

$$\begin{aligned}A_t \max_{\lambda\in\Lambda} \Phi(\lambda, w_{t+1}^{ag}) &\leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 + \alpha_t \lambda_t^{\star,\intercal} f(w_\star) - \frac{1}{2}\|w_{t+1} - w_t\|^2 \\ &+ A_{t-1}\lambda_t^{\star,\intercal} f(w_t^{ag}) + \frac{LA_t}{2}\|w_{t+1}^{ag} - w_t^{md}\|^2 - A_t R(\lambda_t^\star)\end{aligned} \tag{18}$$

By noticing :

$$w_{t+1}^{ag} - w_t^{md} = \frac{\alpha_t}{A_t}\Big(w_{t+1} - w_t\Big)$$

We have after rearranging with f and Φ and decomposing $A_t R(\lambda_t^\star) = (A_{t-1} + \alpha_t)R(\lambda_t^\star)$:

$$\begin{aligned}A_t \max_{\lambda\in\Lambda} \Phi(\lambda, w_{t+1}^{ag}) &\leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 + \left(\frac{L\alpha_t^2}{2A_t} - \frac{1}{2}\right)\|w_{t+1} - w_t\|^2 \\ &+ A_{t-1}\max_{\lambda\in\Lambda} \Phi(\lambda, w_t^{ag}) + \alpha_t \Phi(\lambda_t^\star, w_\star)\end{aligned} \tag{19}$$

By telescopic sum, we have

$$\begin{aligned}A_T \max_{\lambda\in\Lambda} \Phi(\lambda, w_{T+1}^{ag}) &\leq \frac{1}{2}\|w_\star - w_{\text{init}}\|_2^2 + \sum_{t=1}^{T}\left(\frac{L\alpha_t^2}{A_t} - \frac{1}{2}\right)\|w_{t+1} - w_t\|^2 \\ &+ A_0 \max_{\lambda\in\Lambda} \Phi(\lambda, w_{\text{init}}) + \sum_{t=1}^{T}\alpha_t \Phi(\lambda_t^\star, w_\star) \\ &\leq \frac{1}{2}\|w_\star - w_{\text{init}}\|_2^2 + A_0 \max_{\lambda\in\Lambda} \Phi(\lambda, w_{\text{init}}) \\ &+ \sum_{t=1}^{T}\left(\frac{L\alpha_t^2}{2A_t} - \frac{1}{2}\right)\|w_{t+1} - w_t\|^2 + A_t \max_{\lambda\in\Lambda} \Phi(\lambda, w_\star)\end{aligned} \tag{20}$$

From Lemma B.1, the quadratic term cancels. □

D.2 WITH NOISES

We extend the previous result with a precision ϵ_t.

$$\begin{cases} w_t^{md} & = \frac{A_{t-1}}{A_t} w_t^{ag} + \frac{\alpha_t}{A_t} w_t \\ & \lambda_t^\star \quad \text{Fixed point from Theorem 2.1} \\ w_{t+1} & = w_t - \alpha_t \nabla f(w_t^{md}) \lambda_t^\star \\ w_{t+1}^{ag} & = \frac{A_{t-1}}{A_t} w_t^{ag} + \frac{\alpha_t}{A_t} w_{t+1} \end{cases} \tag{21}$$

with the following property Equation (6)

$$\max_{\lambda \in \Lambda} \Phi(\lambda, w_{t+1}^{ag}) - \Phi(\lambda_{t+1}, w_{t+1}^{ag}) \leq \epsilon_t + \frac{L}{2} \|w_{t+1}^{ag} - w_t^{md}\|_2^2.$$

Proof. At iteration t, we have $w_{t+1} = w_t - \alpha_t \nabla f(w_t^{md}) \lambda_{t+1}$, therefore

$$\begin{aligned} \alpha_t \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_{t+1} - w_\star \rangle &= \langle w_t - w_{t+1}, w_{t+1} - w_\star \rangle \\ &= \frac{1}{2} \Big(\|w_\star - w_t\|_2^2 - \|w_\star - w_{t+1}\|_2^2 - \|w_{t+1} - w_t\|_2^2 \Big) \end{aligned}$$

where we use the identity $\langle a, b \rangle = \frac{1}{2} \left(\|a+b\|_2^2 - \|a\|_2^2 - \|b\|_2^2 \right)$. Then, we rearrange the term into

$$\begin{aligned} \alpha_t \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_t^{md} - w_\star \rangle = \frac{1}{2} \|w_\star - w_t\|_2^2 - \frac{1}{2} \|w_\star - w_{t+1}\|_2^2 \\ + \alpha_t \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_t^{md} - w_{t+1} \rangle - \frac{1}{2} \|w_{t+1} - w_t\|_2^2 \end{aligned}$$

We apply the convexity of $\lambda_{t+1} f$,

$$\begin{aligned} \lambda_{t+1} f(w_t^{md}) - \lambda_{t+1} f(w_\star) &\leq \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_t^{md} - w_\star \rangle \\ \alpha_t \Big(\lambda_{t+1} f(w_t^{md}) + \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_{t+1} - w_t^{md} \rangle \Big) &\leq \frac{1}{2} \|w_\star - w_t\|_2^2 - \frac{1}{2} \|w_\star - w_{t+1}\|_2^2 \\ &\quad + \alpha_t \lambda_{t+1} f(w_\star) - \frac{1}{2} \|w_{t+1} - w_t\|_2^2 \end{aligned} \tag{22}$$

On the left hand side, it is related to $\lambda_{t+1} f(w_{t+1}^{ag})$. We use the L-smoothness of $\lambda_{t+1} \cdot f$,

$$\begin{aligned} \Phi(\lambda_{t+1}, w_{t+1}^{ag}) &= \lambda_{t+1} f(w_{t+1}^{ag}) - R(\lambda_{t+1}) \\ &\leq \lambda_{t+1} f(w_t^{md}) + \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_{t+1}^{ag} - w_t^{md} \rangle \\ &\quad + \frac{L}{2} \|w_{t+1}^{ag} - w_t^{md}\|^2 - R(\lambda_{t+1}) \end{aligned}$$

From the Assumption in Equation (6),

$$\begin{aligned} \max_{\lambda \in \Lambda} \Phi(\lambda, w_{t+1}^{ag}) \leq \lambda_{t+1} f(w_t^{md}) + \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_{t+1}^{ag} - w_t^{md} \rangle \\ + L \|w_{t+1}^{ag} - w_t^{md}\|^2 + \epsilon_t - R(\lambda_{t+1}) \end{aligned} \tag{23}$$

Now we make the following decomposition for the scalar product :

$$\begin{aligned} \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_{t+1}^{ag} - w_t^{md} \rangle = \frac{\alpha_t}{A_t} \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_{t+1} - w_t^{md} \rangle \\ + \frac{A_{t-1}}{A_t} \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_t^{ag} - w_t^{md} \rangle \end{aligned} \tag{24}$$

Therefore, Equation (23) scaled by A_t becomes

$$\begin{aligned} A_t \max_{\lambda \in \Delta} \Phi(\lambda, w_{t+1}^{ag}) \leq \alpha_t \left(\lambda_{t+1} f(w_t^{md}) + \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_{t+1} - w_t^{md} \rangle \right) \\ + A_{t-1} \left(\lambda_{t+1} f(w_t^{md}) + \langle \nabla f(w_t^{md}) \lambda_{t+1}, w_t^{ag} - w_t^{md} \rangle \right) \\ + L A_t \|w_{t+1}^{ag} - w_t^{md}\|^2 + A_t \epsilon_t - A_t R(\lambda_{t+1}) \end{aligned} \tag{25}$$

The first line in Equation (25) appears in Equation (22) from Gradient descent.

$$\alpha_t\Big(\lambda_{t+1}f(w_t^{md}) + \langle \nabla f(w_t^{md})\lambda_{t+1}, w_{t+1} - w_t^{md}\rangle\Big) \leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 + \alpha_t\lambda_{t+1}f(w_\star) - \frac{1}{2}\|w_{t+1} - w_t\|^2$$

The second term in Equation (25) is related to $\lambda_{t+1}f(w_t^{ag})$. By the convexity of $\lambda_{t+1}f$

$$A_{t-1}\left(\lambda_{t+1}f(w_t^{md}) + \langle \nabla f(w_t^{md})\lambda_{t+1}, w_t^{ag} - w_t^{md}\rangle\right) \leq A_{t-1}\lambda_{t+1}f(w_t^{ag})$$

Using both simplifications, Equation (25) becomes

$$\begin{aligned} A_t \max_{\lambda\in\Delta} \Phi(\lambda, w_{t+1}^{ag}) \leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 + \alpha_t\lambda_{t+1}f(w_\star) - \frac{1}{2}\|w_{t+1} - w_t\|^2 \\ + A_{t-1}\lambda_{t+1}f(w_t^{ag}) + LA_t\|w_{t+1}^{ag} - w_t^{md}\|^2 + A_t\epsilon_t - A_tR(\lambda_{t+1}) \end{aligned} \tag{26}$$

By noticing :

$$w_{t+1}^{ag} - w_t^{md} = \frac{\alpha_t}{A_t}\Big(w_{t+1} - w_t\Big)$$
$$A_tR(\lambda_{t+1}) = (A_{t-1} + \alpha_t)R(\lambda_{t+1})$$

We have after rearranging :

$$\begin{aligned} A_t \max_{\lambda\in\Delta} \Phi(\lambda, w_{t+1}^{ag}) &\leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 + \left(\frac{L\alpha_t^2}{A_t} - \frac{1}{2}\right)\|w_{t+1} - w_t\|^2 \\ &+ A_{t-1}\Big(\lambda_{t+1}f(w_t^{ag}) - R(\lambda_{t+1})\Big) + \alpha_t\Big(\lambda_{t+1}f(w_\star) - R(\lambda_{t+1})\Big) + A_t\epsilon_t \\ &\leq \frac{1}{2}\|w_\star - w_t\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 + \left(\frac{L\alpha_t^2}{A_t} - \frac{1}{2}\right)\|w_{t+1} - w_t\|^2 \\ &+ A_{t-1}\max_{\lambda\in\Lambda}\Phi(\lambda, w_t^{ag}) + \alpha_t\Phi(\lambda_{t+1}, w_\star) + A_t\epsilon_t \end{aligned} \tag{27}$$

From Lemma B.1, the quadratic terms cancel and we conclude by applying telescopic sum as before. □

E CROSS-EVALUATION GAP

We are demonstrating Theorem 3.1 which shows that $\forall\lambda, \mu \in \Lambda$

$$\Phi(\mu, w_{t+1}^{ag}(\lambda)) - \Phi(\lambda, w_{t+1}^{ag}(\lambda)) \leq \langle \nabla S_t(\lambda), \lambda - \mu\rangle + R(\lambda) - R(\mu) + \frac{L}{2}\|w_{t+1}^{ag}(\lambda) - w_t^{md}\|_2^2$$

Proof. We recall that $V_t(\lambda) := \nabla_w\Phi(\lambda, w_t^{md})$. By the definition of the Nesterov aggregate sequence, we have:

$$\begin{aligned} w_{t+1}^{ag}(\lambda) &= w_t^{md} + \frac{\alpha_t}{A_t}(w_{t+1}(\lambda) - w_t) \\ w_{t+1}^{ag}(\lambda) - w_t^{md} &= -\frac{\alpha_t^2}{A_t}V_t(\lambda) = -\frac{\alpha_t^2}{A_t}\nabla f(w_t^{md})\lambda \end{aligned} \tag{28}$$

Using the L-smoothness of Φ with respect to the primal variable w ,we upper-bound the function value for any chosen adversary $\mu \in \Lambda$:

$$\Phi(\mu, w_{t+1}^{ag}(\lambda)) \leq \Phi(\mu, w_t^{md}) + \langle V_t(\mu), w_{t+1}^{ag}(\lambda) - w_t^{md}\rangle + \frac{L}{2}||w_{t+1}^{ag}(\lambda) - w_t^{md}||_2^2$$

Simultaneously, using the convexity of Φ with respect to w for the dual variable $\lambda \in \Lambda$, we lower-bound its value:

$$\Phi(\lambda, w_{t+1}^{ag}(\lambda)) \geq \Phi(\lambda, w_t^{md}) + \langle V_t(\lambda), w_{t+1}^{ag}(\lambda) - w_t^{md}\rangle$$

Subtracting the convexity inequality from the smoothness inequality isolates the cross-evaluation gap:

$$\Phi(\mu, w_{t+1}^{ag}(\lambda)) - \Phi(\lambda, w_{t+1}^{ag}(\lambda)) \leq \Phi(\mu, w_t^{md}) - \Phi(\lambda, w_t^{md}) + \langle V_t(\mu) - V_t(\lambda), w_{t+1}^{ag}(\lambda) - w_t^{md} \rangle + \frac{L}{2} ||w_{t+1}^{ag}(\lambda) - w_t^{md}||_2^2 \tag{29}$$

We now substitute Equation (28) into the inner product term:

$$\begin{aligned} \langle V_t(\mu) - V_t(\lambda), w_{t+1}^{ag}(\lambda) - w_t^{md} \rangle &= -\frac{\alpha_t^2}{A_t} \langle \nabla f(w_t^{md})(\mu - \lambda), \nabla f(w_t^{md}) \lambda \rangle \\ &= \frac{\alpha_t^2}{A_t} \langle \lambda - \mu, \nabla f(w_t^{md})^\intercal \nabla f(w_t^{md}) \lambda \rangle \end{aligned}$$

Now we decompose Φ,

$$\Phi(\mu, w_t^{md}) - \Phi(\lambda, w_t^{md}) = (\mu - \lambda) \cdot f(w_t^{md}) + R(\lambda) - R(\mu)$$

Substituting these two lines into Equation (29), we obtain

$$\begin{aligned} \Phi(\mu, w_{t+1}^{ag}(\lambda)) - \Phi(\lambda, w_{t+1}^{ag}(\lambda)) \leq \langle \frac{\alpha_t^2}{A_t} \nabla f(w_t^{md})^\intercal \nabla f(w_t^{md}) \lambda - f(w_t^{md}), \lambda - \mu \rangle \\ + \frac{L}{2} ||w_{t+1}^{ag}(\lambda) - w_t^{md}||_2^2 + R(\lambda) - R(\mu) \end{aligned}$$

we conclude by noticing from Equation (5) that

$$\nabla S_t(\lambda) = \frac{\alpha_t^2}{A_t} \nabla f(w_t^{md})^\intercal \nabla f(w_t^{md}) \lambda - f(w_t^{md}).$$

□

E.1 General upper bound for smooth concave-convex function

In this subsection, we suppose that Φ is a general smooth concave-convex function, the general form is the Moreau envelop

$$\begin{aligned} V_t(\lambda) &:= \nabla_w \Phi(\lambda, w_t^{md}) \\ h_t(\lambda) &:= \frac{\alpha_t^2}{2A_t} \|V_t(\lambda)\|_2^2 - \Phi(\lambda, w_t^{md}) \end{aligned}$$

We will show the new upper bound in the general setting

Theorem E.1. *For any $\lambda, \mu \in \Lambda$, the cross-evaluation gap is strictly upper-bounded by:*

$$\Phi(\mu, g_t(\lambda)) - \Phi(\lambda, g_t(\lambda)) \leq \Big[h_t(\lambda) - h_t(\mu) \Big] + \frac{\alpha_t^2}{2A_t} \Big\| V_t(\mu) - V_t(\lambda) \Big\|_2^2 + \frac{L\alpha_t^4}{2A_t^2} \Big\| V_t(\lambda) \Big\|_2^2$$

where h_t refers to Equation (5) *and $g_t(\lambda) := w_{t+1}^{ag}(\lambda)$.*

Proof. By the definition of the Nesterov aggregate sequence, we have:

$$\begin{aligned} w_{t+1}^{ag}(\lambda) &= w_t^{md} + \frac{\alpha_t}{A_t}(w_{t+1}(\lambda) - w_t) \\ w_{t+1}^{ag}(\lambda) - w_t^{md} &= -\frac{\alpha_t^2}{A_t} V_t(\lambda) \end{aligned} \tag{30}$$

Using the L-smoothness of Φ with respect to the primal variable w, we upper-bound the function value for any chosen adversary $\mu \in \Lambda$:

$$\Phi(\mu, w_{t+1}^{ag}(\lambda)) \leq \Phi(\mu, w_t^{md}) + \langle V_t(\mu), w_{t+1}^{ag}(\lambda) - w_t^{md} \rangle + \frac{L}{2} ||w_{t+1}^{ag}(\lambda) - w_t^{md}||_2^2$$

Simultaneously, using the convexity of Φ with respect to w for the dual variable $\lambda \in \Lambda$, we lower-bound its value:

$$\Phi(\lambda, w_{t+1}^{ag}(\lambda)) \geq \Phi(\lambda, w_t^{md}) + \langle V_t(\lambda), w_{t+1}^{ag}(\lambda) - w_t^{md} \rangle$$

Subtracting the convexity inequality from the smoothness inequality isolates the cross-evaluation gap:

$$\Phi(\mu, w_{t+1}^{ag}(\lambda)) - \Phi(\lambda, w_{t+1}^{ag}(\lambda)) \leq \Phi(\mu, w_t^{md}) - \Phi(\lambda, w_t^{md}) + \langle V_t(\mu) - V_t(\lambda), w_{t+1}^{ag}(\lambda) - w_t^{md} \rangle + \frac{L}{2}||w_{t+1}^{ag}(\lambda) - w_t^{md}||_2^2$$

We now substitute Equation (30) into the inner product term:

$$\begin{aligned}\langle V_t(\mu) - V_t(\lambda), w_{t+1}^{ag}(\lambda) - w_t^{md} \rangle &= -\frac{\alpha_t^2}{A_t}\langle V_t(\mu) - V_t(\lambda), V_t(\lambda) \rangle \\ &= -\frac{\alpha_t^2}{A_t}\langle V_t(\mu), V_t(\lambda) \rangle + \frac{\alpha_t^2}{A_t}||V_t(\lambda)||_2^2\end{aligned}$$

By employing the standard algebraic identity $-2\langle a, b \rangle = ||a - b||_2^2 - ||a||_2^2 - ||b||_2^2$, we rewrite the inner product dynamically:

$$-\frac{\alpha_t^2}{A_t}\langle V_t(\mu), V_t(\lambda) \rangle = \frac{\alpha_t^2}{2A_t}||V_t(\mu) - V_t(\lambda)||_2^2 - \frac{\alpha_t^2}{2A_t}||V_t(\mu)||_2^2 - \frac{\alpha_t^2}{2A_t}||V_t(\lambda)||_2^2$$

Re-adding the $\frac{\alpha_t^2}{A_t}||V_t(\lambda)||_2^2$ expansion component, the total cross-term equates strictly to:

$$\langle V_t(\mu) - V_t(\lambda), w_{t+1}^{ag}(\lambda) - w_t^{md} \rangle = \frac{\alpha_t^2}{2A_t}||V_t(\mu) - V_t(\lambda)||_2^2 - \frac{\alpha_t^2}{2A_t}||V_t(\mu)||_2^2 + \frac{\alpha_t^2}{2A_t}||V_t(\lambda)||_2^2$$

Substitute this resulting identity back into the main cross-evaluation gap inequality and group the variables by λ and μ:

$$\Phi(\mu, w_{t+1}^{ag}(\lambda)) - \Phi(\lambda, w_{t+1}^{ag}(\lambda)) \leq \left(\frac{\alpha_t^2}{2A_t}||V_t(\lambda)||_2^2 - \Phi(\lambda, w_t^{md})\right) - \left(\frac{\alpha_t^2}{2A_t}||V_t(\mu)||_2^2 - \Phi(\mu, w_t^{md})\right) + \frac{\alpha_t^2}{2A_t}||V_t(\mu) - V_t(\lambda)||_2^2 + \frac{L}{2}||w_{t+1}^{ag}(\lambda) - w_t^{md}||_2^2$$

We recognize by Equation (5) that the first two terms correspond to Moreau Envelope Proxy. Then, we address the terminal L-smoothness bounding term $w_{t+1}^{ag}(\lambda) - w_t^{md} = \frac{\alpha_t}{A_t}(w_{t+1}(\lambda) - w_t)$. Therefore:

$$\frac{L}{2}||w_{t+1}^{ag}(\lambda) - w_t^{md}||_2^2 = \frac{L\alpha_t^2}{2A_t^2}||w_{t+1}(\lambda) - w_t||_2^2 = \frac{L\alpha_t^4}{2A_t^2}\|V_t(\lambda)\|_2^2$$

Combining all simplified blocks yields the desired bounding theorem exactly. □

F KKT CONDITION FOR QUADRATIC CASE

We recall Lemma 4.1.

Lemma F.1. *Suppose an interior-point algorithm terminates at a strictly feasible output $\mu_k \in$ relint(Δ_J), returning a barrier parameter $1/\tau_k \leq \epsilon_\tau$ and dual multipliers $\nu_k \in \mathbb{R}$ and $z_k \in \mathbb{R}_+^J$ that satisfy the following inexact Karush-Kuhn-Tucker (KKT) conditions:*

- ***Approximate Stationarity*** $||\nabla h_t(\mu_k) - \nu_k 1 - z_k||_\infty \leq \epsilon_g$
- ***Perturbed Complementary Slackness*** $\mu_{k,j} z_{k,j} = 1/\tau_k$ *for all* $j \in \{1, ..., J\}$

Then, the Frank-Wolfe duality gap at μ_k is strictly bounded by:

$$\max_{\mu \in \Lambda} \langle \nabla h_t(\mu_k), \mu_k - \mu \rangle \leq J\epsilon_\tau + 2\epsilon_g$$

Proof. Because h_t is convex, the true suboptimality is upper-bounded by the Frank-Wolfe gap at μ_k:

$$\max_{\mu \in \Lambda} \langle \nabla h_t(\mu_k), \mu_k - \mu \rangle = \langle \nabla h_t(\mu_k), \mu_k \rangle - \min_{\mu \in \Delta_J} \langle \nabla h_t(\mu_k), \mu \rangle$$

Let $r_k = \nabla h_t(\mu_k) - \nu_k 1 - z_k$ denote the stationarity residual, by assumption we have $||r_k||_\infty \leq \epsilon_g$. Substituting $\nabla h_t(\mu_k) = \nu_k 1 + z_k + r_k$ yields

$$\begin{aligned}\langle \nabla h_t(\mu_k), \mu_k \rangle &= \nu_k \langle \mathbf{1}, \mu_k \rangle + \langle z_k, \mu_k \rangle + \langle r_k, \mu_k \rangle \\ &= \nu_k + J/\tau_k + \langle r_k, \mu_k \rangle\end{aligned}$$

Then, we notice that

$$\begin{aligned}\min_{y \in \Delta_J} \langle \nabla h_t(\mu_k), \mu \rangle &= \min_{y \in \Delta_J} \langle \nu_k 1 + z_k + r_k, \mu \rangle = \nu_k + \min_j (z_{k,j} + r_{k,j}) \\ \max_{\mu \in \Lambda} \langle \nabla h_t(\mu_k), \mu_k - \mu \rangle &= J/\tau_k + \langle r_k, \mu_k \rangle - \min_j (z_{k,j} + r_{k,j})\end{aligned}$$

Since $||r_k||_\infty \leq \epsilon_g$, by Holder inequality:

$$\begin{aligned}\langle r_k, \mu_k \rangle &\leq ||r_k||_\infty ||\mu_k||_1 \leq \epsilon_g \\ \forall 1 \leq j \leq J, \quad z_{k,j} + r_{k,j} > r_{k,j} &\geq -||r_k||_\infty \geq -\epsilon_g \\ \implies -\min_j (z_{k,j} + r_{k,j}) &\leq \epsilon_g\end{aligned}$$

Since the algorithm ensures $1/\tau_k \leq \epsilon_\tau$, we conclude:

$$g_{FW}(\mu_k) \leq J\epsilon_\tau + 2\epsilon_g.$$

□

G PATH-FOLLOWING ALGORITHM

We rewrite the path-following algorithm Algorithm 2 from (Nesterov & Nemirovski, 1994). From

Algorithm 2 Centered path-following method

1: **Initialize:** Set $\mu_0 \in \arg\min R + B$ and $\tau_0 = 1$.
2: **Phase I:** Initial Centering (Damped Step)
3: **while** $\delta(\mu_k, F_1) > 1/4$ **do**
4: Compute Newton Step $\Delta\mu_k$ for (μ_k, F_1) with Equation (34)
5: Update: $\mu_{k+1} = \mu_k + \frac{1}{1+\delta(\mu_k, F_1)} \Delta\mu_k$, $k \leftarrow k+1$
6: **end while**
7: **Phase II:** Short-Step Path-Following(Normal Full Step)
8: **while** Target accuracy ϵ_t is not reached **do**
9: Update Penalty: $\tau_{k+1} = \tau_k \times \left(1 + \frac{1}{15\sqrt{J}}\right)$
10: Compute Newton Step $\Delta\mu_k$ for $(\mu_k, F_{\tau_{k+1}})$ with Equation (34)
11: Update: $\mu_{k+1} = \mu_k + \Delta\mu_k$ (, $k \leftarrow k+1$
12: **end while**
13: **return** μ_k

Nesterov framework, we show in the following Theorem about equality-constrained Newton method.

Theorem G.1. *Consider $F : \mathbb{R}^J_{++} \to \mathbb{R}$ strictly convex and 1-self-concordant, let μ_k be a strictly feasible solution and $(\Delta\mu_k, s_k)$ be defined as the solution of the KKT block system:*

$$\begin{bmatrix} \nabla^2 F(\mu_k) & \mathbf{1} \\ \mathbf{1}^\intercal & 0 \end{bmatrix} \begin{bmatrix} \Delta\mu_k \\ s_k \end{bmatrix} = \begin{bmatrix} -\nabla F(\mu_k) \\ 0 \end{bmatrix} \tag{31}$$

Let $\omega(\lambda) = \lambda - \ln(1+\lambda)$ and $\lambda_k = \delta(\mu_k, F)$, then the following properties hold for the constrained updates.

1. ***Damped Newton Step:*** *Taking the step $\mu_{k+1} = \mu_k + \frac{1}{1+\lambda_k} \Delta\mu_k$ maintains strict feasibility, and the objective strictly decreases by:*

$$F(\mu_{k+1}) \leq F(\mu_k) - \omega(\lambda_k) \tag{32}$$

2. ***Full Newton Step:*** *If $\delta(\mu_k, F) < 1$, taking a full Newton step $\mu_{k+1} = \mu_k + \Delta\mu_k$ maintains strict feasibility, converges quadratically, and the new Newton decrement is bounded by:*

$$\delta(\mu_{k+1}, F) \leq \left(\frac{\delta(\mu_k, F)}{1 - \delta(\mu_k, F)}\right)^2 \tag{33}$$

The proof can be found found in Theorem 5.1.15, 5.2.2 in Nesterov (2018) and Section 10.2.3 in Boyd & Vandenberghe (2004).

If $\nabla^2 F_\tau > 0$, we can efficiently calculate the step solution $(\Delta\mu, s)$ via block elimination:

$$\begin{aligned} s &\leftarrow -\frac{\mathbf{1}^\intercal(\nabla^2 F_\tau(\mu))^{-1}\nabla F_\tau(\mu)}{\mathbf{1}^\intercal(\nabla^2 F_\tau(\mu))^{-1}\mathbf{1}}, \\ \Delta\mu &\leftarrow -(\nabla^2 F_\tau(\mu))^{-1}(\nabla F_\tau(\mu) + s\mathbf{1}). \end{aligned} \tag{34}$$

G.1 Cross-evaluation gap bound

Before the main result, we cite a property of self-concordant-barrier from (Nesterov & Nemirovski, 1994).

Lemma G.2. *If $B_\Lambda : \mathrm{relint}(\Lambda) \to \mathbb{R}$ is a ν-self-concordant barrier for the compact convex set Λ. Then for all $\mu \in \mathrm{relint}(\Lambda)$ and $\lambda \in \Lambda$:*

$$\begin{aligned} \langle \nabla B_\Lambda(\mu), \lambda - \mu \rangle &\leq \nu \\ \|\nabla B_\Lambda(\mu)\|^*_{\nabla^2 B_\Lambda(\mu)} &\leq \sqrt{\nu} \end{aligned}$$

We recall the main Theorem to prove.

Theorem G.3 (Inner problem Upper bound via KKT block)**.** *Under Assumption 1.1, we consider a fixed $\tau \geq 1$ and $\mu_k \in \mathrm{relint}(\Lambda)$. If the KKT block on on (μ_k, F_τ) has a feasible solution $(\Delta\mu_k, s_k)$ and there exists $\beta \in [0, 1]$,*

$$\delta(\mu_k, F_\tau) \leq \beta^2,$$

then for all $\lambda \in \Lambda$

$$\epsilon_t(\mu_k, \lambda) \leq \frac{\nu}{\tau} + \frac{\sqrt{\nu}}{\tau}\left(\frac{\beta}{1-\beta}\right) + \frac{\omega_*(\beta)}{\tau} + \frac{\beta}{1-\beta}\frac{2G\alpha_t}{\sqrt{\tau A_t}} \tag{35}$$

where $\omega_*(\beta) := -\beta - \ln(1-\beta)$.

Proof. We recall that

$$\begin{aligned} \epsilon_t(\lambda, \mu) &= \langle \nabla S_t(\lambda), \lambda - \mu \rangle + R(\lambda) - R(\mu) \\ F_\tau &= \tau(S_t + R) + B \end{aligned}$$

We can rewrite for $\lambda, \mu^\star \in \Lambda$

$$\begin{aligned} \epsilon_t(\mu_k, \lambda) &= \langle \nabla S_t(\mu_k), \mu_k - \lambda \rangle + R(\mu_k) - R(\lambda) \\ &= \langle \nabla S_t(\mu_k) - \nabla S_t(\mu^\star), \mu_k - \lambda \rangle + \langle \nabla S_t(\mu^\star), \mu_k - \lambda \rangle + R(\mu_k) - R(\lambda). \end{aligned} \tag{36}$$

Now we fix $\mu^\star := \arg\min_{\mu \in \Lambda} F_\tau$, since $\mu^\star, \mu_k \in \mathrm{relint}(\Lambda)$, for all $\lambda \in \Lambda, \mu_k - \lambda$ is a feasible position and therefore $\tau > 0, \lambda \in \Lambda$:

$$\begin{aligned} & \langle \nabla F_\tau(\mu^\star), \mu_k - \lambda \rangle = 0 \\ \implies \quad & \langle \nabla S_t(\mu^\star), \mu_k - \lambda \rangle = -\langle \nabla R(\mu^\star), \mu_k - \lambda \rangle - \frac{1}{\tau}\langle \nabla B(\mu^\star), \mu_k - \lambda \rangle \end{aligned}$$

We notice that

$$R(\mu_k) - R(\lambda) - \langle \nabla R(\mu^\star), \mu_k - \lambda \rangle = D_R(\mu_k, \mu^\star) - D_R(\lambda, \mu^\star)$$

Therefore, Equation (36) can be rewritten as

$$\begin{aligned}\epsilon_t(\mu_k,\lambda) &= \langle \nabla S_t(\mu_k) - \nabla S_t(\mu^\star), \mu_k - \lambda\rangle - \frac{1}{\tau}\langle \nabla B(\mu^\star), \mu_k - \lambda\rangle + D_R(\mu_k,\mu^\star) - D_R(\lambda,\mu^\star) \\ &\leq \langle \nabla S_t(\mu_k) - \nabla S_t(\mu^\star), \mu_k - \lambda\rangle - \frac{1}{\tau}\langle \nabla B(\mu^\star), \mu_k - \lambda\rangle + D_R(\mu_k,\mu^\star).\end{aligned} \tag{37}$$

For the first quadratic term, because $\nabla S_t(\mu_k) - \nabla S_t(\mu^\star) = H_Q(\mu_k - \mu^\star)$ where $H_Q = \frac{\alpha_t^2}{A_t}\nabla f(w_t^{md})^\intercal \nabla f(w_t^{md})$.

We use the generalized Cauchy-Schwarz inequality. Since $\tau H_Q \preceq \nabla^2 F_\tau$, we have

$$\|\mu_k - \mu^\star\|_{H_Q} \leq \frac{1}{\sqrt{\tau}}\|\mu_k - \mu^\star\|_{\nabla^2 F_\tau(\mu^\star)} \leq \frac{1}{\sqrt{\tau}}\frac{\beta}{1-\beta}.$$

And from Assumption 1.1, $\|\nabla f(w_t^{md})\lambda\|_2 \leq G$,

$$\|\mu_k - \lambda\|_{H_Q} = \frac{\alpha_t}{\sqrt{A_t}}\|\nabla f(w_t^{md})(\mu_k - \lambda)\|_2 \leq \frac{2G\alpha_t}{\sqrt{A_t}}$$

Therefore,

$$\langle H_Q(\mu_k - \mu^\star), \mu_k - \lambda\rangle \leq \frac{\beta}{1-\beta}\frac{2G\alpha_t}{\sqrt{\tau A_t}}$$

For the second Barrier term, from Lemma G.2, we notice that for any arbitrary $\mu^\star \in \operatorname{relint}(\Lambda), \lambda \in \Lambda$,

$$\langle \nabla B(\mu^\star), \lambda - \mu^\star\rangle \leq \nu$$

For the remaining difference $-\frac{1}{\tau}\langle \nabla B(\mu^\star), \mu_k - \mu^\star\rangle$, by self-concordance theory Nesterov & Nemirovski (1994), Theorem 5.1.8 (Equation 5.1.28), if $\delta(\mu_k, F_\tau) \leq \beta < 1$ the distance to the exact optimum $\mu^\star$ strictly is bounded by

$$\|\mu_k - \mu^\star\|_{\nabla^2 F_\tau(\mu^\star)} \leq \frac{\beta}{1-\beta},$$

and the dual norm of $\|\nabla B(\mu^\star)\|_* \leq \sqrt{\nu}$ (Lemma G.2), therefore

$$-\frac{1}{\tau}\langle \nabla B(\mu^\star), \mu_k - \lambda\rangle \leq \frac{\nu}{\tau} + \frac{\sqrt{\nu}}{\tau}\frac{\beta}{1-\beta}$$

For the last term, we notice , $D_{F_\tau}(\mu_k, \mu^\star) \leq \omega_*(\beta)$ from (Nesterov & Nemirovski, 1994), therefore

$$D_R(\mu_k,\mu^\star) \leq \frac{1}{\tau}D_{F_\tau}(\mu_k,\mu^\star) \leq \frac{\omega_*(\beta)}{\tau}$$

□

G.2 Phase I complexity

Theorem G.4 (Phase I Complexity)**.** *Let $\tau_0 = 1$. Under Assumptions 1.1 and 4.1, starting from $\mu_0 = \frac{1}{J}\mathbf{1}$, the Damped Newton method strictly reaches the short-step neighborhood $\mathcal{N}(\beta) = \{\mu \in \operatorname{relint}(\Delta_J) \mid \delta(\mu, F_1) \leq \beta\}$ with $\beta \leq 1$ in a bounded number of iterations:*

$$N_{\textit{Phase I}} \leq \frac{1}{\omega(1/4)}\left(\frac{G^2\alpha_t^2}{2A_t} + 2G\|w_t^{md} - w_\star\|_2 + S_\star\right)$$

Proof. First we recall that $S_t(\mu)$ is quadratic and F_1 is strictly 1-self-concordant. By Theorem G.1, we know that for the each damped Newton step $\mu_{\text{new}} = \mu + \frac{1}{1+\delta(\mu)}\Delta\mu$, we have

$$F_1(\mu) - F_1(\mu_{\text{new}}) \geq \omega(\lambda_k) \geq \omega(1/4)$$

where $\omega(\beta) = \beta - \ln(1+\beta) > 0$. Consequently, the maximum number of damped iterations required to enter the neighborhood is strictly bounded by the total initial suboptimality gap divided by this minimum step decrease:

$$N_{\text{Phase I}} \leq \frac{F_1(\mu_0) - F_1(\mu_1^\star)}{\omega(1/4)}$$

where $\mu_1^\star = \arg\min_{\mu \in \Delta_J} F_1(\mu)$. We notice that,

$$F_1(\mu_0) - F_1(\mu_1^\star) = \Big[S_t(\mu_0) - S_t(\mu_1^\star)\Big] + \Big[(R+B)(\mu_0) - (R+B)(\mu_1^*)\Big].$$

Since $\mu_0 \in \arg\min R + B$, we have

$$(R+B)(\mu_0) - (R+B)(\mu_1^\star) \leq 0$$
$$F_1(\mu_0) - F_1(\mu_1^\star) \leq S_t(\mu_0) - S_t(\mu_1^\star).$$

We recall that :

$$S_t(\lambda) = \frac{\alpha_t^2}{2A_t}\|\nabla f(w_t^{md})\lambda\|_2^2 - \lambda^\intercal f(w_t^{md}) \geq -\lambda^\intercal f(w_t^{md})$$
$$S_t(\mu_0) - S_t(\mu_1^\star) \leq \frac{\alpha_t^2}{2A_t}\|\nabla f(w_t^{md})\mu_0\|_2^2 + (\mu_1^\star - \mu_0)^\intercal f(w_t^{md}).$$

We apply triangular inequality and obtain that

$$(\mu_1^\star - \mu_0)^\intercal f(w_t^{md}) \leq 2\max_{j,k}|f_j(w_t^{md}) - f_k(w_t^{md})| \leq 2G\|w_t^{md} - w_\star\|_2 + S_\star$$

and we conclude that

$$\frac{\alpha_t^2}{2A_t}\|\nabla f(w_t^{md})\mu_0\|_2^2 \leq \frac{G^2\alpha_t^2}{2A_t}$$
$$F_1(\mu_0) - F_1(\mu_1^\star) \leq \frac{G^2\alpha_t^2}{2A_t} + 2G\|w_t^{md} - w_\star\|_2 + S_\star.$$

□

G.3 Phase II complexity

We recall that $F_{\tau_k}(\mu) := \tau_k\Big(S_t(\mu) + R(\mu)\Big) + B(\mu)$ where $B(\mu) = -\sum_{j=1}^J \ln(\mu_j)$.

Theorem G.5 (Phase II Complexity)**.** *Suppose the current iterate $\mu_k \in \text{relint}(\Delta_J)$ satisfies the Newton decrement neighborhood condition*

$$\delta(\mu_k, F_{\tau_k}) \leq 1/4,$$

If we update $\tau_{k+1} = r\tau_k$ with $r = 1 + \frac{1}{15\sqrt{J}}$, then taking exactly a full Newton step guarantees the new iterate strictly remains in the neighborhood: $\delta(\mu_{k+1}, F_{\tau_{k+1}}) \leq \frac{1}{4}$.

Proof. We recall that

$$F_{\tau_{k+1}}(\mu) = r\tau_k\Big(S_t(\mu) + R(\mu)\Big) + B(\mu) = rF_{\tau_k}(\mu) - (r-1)B(\mu)$$

Taking the gradient yields $\nabla F_{\tau_{k+1}}(\mu) = r\nabla F_{\tau_k}(\mu) - (r-1)\nabla B(\mu)$. Furthermore, because $S_t + R$ is convex, we have the two following inequalities:

$$\nabla^2 F_{\tau_{k+1}}(\mu) \succeq \nabla^2 F_{\tau_k}(\mu) \quad \text{and} \quad \nabla^2 F_{\tau_{k+1}}(\mu) \succeq \nabla^2 B(\mu)$$

Let $(\Delta\mu_k, s_k)$ be the solution of the KKT system $(\mu_k, F_{\tau_{k+1}})$ Definition 4.1. The intermediate Newton decrement can be rewritten

$$\begin{aligned}\delta_{\text{shift}}^2 &:= \delta(\mu_k, F_{\tau_{k+1}})^2 \\ &= \Delta\mu_k^\intercal \nabla^2 F_{\tau_{k+1}}(\mu_k)\Delta\mu_k \\ &= -\Delta\mu_k^\intercal \nabla F_{\tau_{k+1}}(\mu_k) = \underbrace{r\left(-\Delta\mu_k^\intercal \nabla F_{\tau_k}(\mu_k)\right)}_{\text{Term A}} + \underbrace{(r-1)\Delta\mu_k^\intercal \nabla B(\mu_k)}_{\text{Term B}}\end{aligned}$$

On a hand, let (ν_k, s_k) be the solution of the KKT system with (μ_k, F_{τ_k}) Definition 4.1. By Cauchy-Schwarz,

$$-\Delta\mu_k^\intercal \nabla F_{\tau_k}(\mu_k) = \Delta\mu_k^\intercal \nabla^2 F_{\tau_k}(\mu_k) v_k \leq \|\Delta\mu_k\|_{\nabla^2 F_{\tau_k}} \|v_k\|_{\nabla^2 F_{\tau_k}}$$

By the assumption that the previous step was centered,

$$\|v_k\|_{\nabla^2 F_{\tau_k}} = \delta(\mu_k, F_{\tau_k}) \leq \beta$$
$$\|\Delta\mu_k\|_{\nabla^2 F_{\tau_k}} \leq \|\Delta\mu_k\|_{\nabla^2 F_{\tau_{k+1}}} = \delta_{\text{shift}}.$$

Combining all the equation, the Term A is smaller than $r\beta\delta_{\text{shift}}$.

On the other hand, we have

$$\Delta\mu_k^\intercal \nabla B(\mu_k) \leq \|\Delta\mu_k\|_{\nabla^2 B} \|\nabla B(\mu_k)\|_B^*$$

Because $\nabla^2 B \preceq \nabla^2 F_{\tau_{k+1}}$, the first term is bounded by δ_{shift}. By the fundamental property of a J-self-concordant barrier (Nesterov (2018) Definition 5.3.1), its local dual norm is globally bounded: $\|\nabla B(\mu_k)\|_B^* \leq \sqrt{J}$. Thus, Term B $\leq (r-1)\sqrt{J}\delta_{\text{shift}}$.

Substituting both terms back into the expansion gives

$$\begin{aligned}\delta_{\text{shift}} &\leq r\beta + (r-1)\sqrt{J} \\ &= \left(1 + \frac{\gamma}{\sqrt{J}}\right)\beta + \gamma \leq \beta + \gamma(1+\beta)\end{aligned}$$

Setting $\beta = 1/4, \gamma = 1/15$, we obtain $\delta_{\text{shift}} \leq 1/3 < 1$. Because $F_{\tau_{k+1}}$ is strictly 1-self-concordant, and the intermediate error is safely < 1, classical single-step Newton convergence (Theorem G.1) applies. Exactly one normal, undamped Newton step drops the decrement quadratically:

$$\delta(\mu_{k+1}, F_{\tau_{k+1}}) \leq \left(\frac{\delta_{\text{shift}}}{1-\delta_{\text{shift}}}\right)^2 \leq \left(\frac{1/3}{1-1/3}\right)^2 = \frac{1}{4} = \beta$$

□

G.4 Overall complexity

We recall the result to prove :

Corollary G.6. *Suppose that $\|w_\star - w_1\|_2 \leq D$. By fixing $\beta = \frac{1}{4}$, $\epsilon_t := \frac{L}{t^4}$, under Assumption 1.1, we have*

$$N_{\textit{Phase I}} + N_{\textit{Phase II}} = \mathcal{O}\left(G + \sqrt{J}\log\left(JGt\right)\frac{G^2}{L} + 2G\sqrt{D^2 + \log(t)} + S_\star\right)$$

Proof. First, we notice that

$$N_{\text{Phase II}} = \mathcal{O}\left(G + \sqrt{J}\log\left(JGt\right)\right)$$

In order to show the upper bound of the Phase I complexity, we recall that Equation (27) implies for any $t \geq 1$

$$\begin{aligned}A_t \max_{\lambda\in\Delta} \Phi(\lambda, w_{t+1}^{ag}) &\leq \frac{1}{2}\|w_\star - w_1\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 + \sum_{\tau=1}^{t} \alpha_t \Phi(\lambda_{t+1}, w_\star) + \sum_{\tau=1}^{t} A_\tau \epsilon_\tau \\ &\leq \frac{1}{2}\|w_\star - w_1\|_2^2 - \frac{1}{2}\|w_\star - w_{t+1}\|_2^2 + A_t \max_{\lambda\in\Delta} \Phi(\lambda, w_\star) + \sum_{\tau=1}^{t} A_\tau \epsilon_\tau\end{aligned} \tag{38}$$

Therefore, since $w_\star$ is the minimum by definition, we deduce that

$$\|w_\star - w_{t+1}\|_2^2 \leq \|w_\star - w_1\|_2^2 + 2\sum_{\tau=1}^{t} A_\tau \epsilon_\tau \leq D^2 + \mathcal{O}(\log(t))$$

We also observe that w_t^{md} is a convex combination of w_t, therefore for all $t \geq 1$

$$\|w_\star - w_t^{md}\|_2 \leq \max_{1 \leq \tau \leq t} \|w_\star - w_\tau\|_2$$

Theorem G.4 implies that

$$N_{\text{Phase I}} = \mathcal{O}\left(\frac{G^2}{L} + 2G\sqrt{D^2 + \log(t)} + S_\star\right)$$

which concludes our proof. □

H SPECTRAHEDRON

The *spectrahedron* is defined as the set of $J \times J$ symmetric PSD matrices with unit trace

$$\mathcal{S}_J := \{\lambda \in \mathbb{S}_+^J \mid \text{Tr}(\lambda) = 1\}.$$

The affine coupling for matrices becomes :

$$\min_{w \in \mathbb{R}^d} \max_{\lambda \in S_J} \text{Tr}(\lambda F(w)) - R(\lambda)$$

where $F : \mathbb{R}^d \to S_J$ is a matrix-convex (Loewner-convex) and $R : \mathbb{R}^J \to \mathbb{R}$ convex.

For the spectrahedron we use the log-determinant barrier:

$$\begin{aligned} B_\Lambda(\lambda) &:= -\log\det(\lambda) \\ F_\tau(\lambda) &:= \tau(S_t(\lambda) + R(\lambda)) - \log\det(\lambda) \end{aligned} \tag{39}$$

This function is strictly J-self-concordant for the PSD cone $\mathbb{S}_+^J$ (Nesterov, 2018).

I SIMULATIONS

Note that data generation and the pre-computation of the smoothness constant L are excluded from the runtime measurements. Because the exact optimal primal value is intractable, we approximate it empirically using the lowest objective value achieved across all runs of our proposed method. **Data generation**

- A true underlying weight vector: $w_{true} \sim \mathcal{N}\left(0, \frac{20}{d} I\right)$
- Group-specific distribution shift: $\mu_j \sim \mathcal{N}\left(0, \frac{1}{d} I\right)$
- Sample generation: $x_{j,i} \sim \mathcal{N}\left(0, \frac{1}{d} I\right) + 0.1\mu_j$
- Label Generation: $y_{j,i} = \text{sign}(\langle x_{j,i}, w_{true}\rangle + \varepsilon_{j,i})$ where $\varepsilon_{j,i} \sim \mathcal{N}(0, 1)$

Gradient computation

$$\nabla f_j(w) = \frac{1}{N} \sum_{i=1}^{N} \frac{-y_{j,i} X_{j,i}}{1 + \exp(y_{j,i} X_{j,i}^\intercal w)}$$

$$L_j := \lambda_{\max}\left(\frac{1}{4N} X_j^\intercal X_j\right)$$

$$L := \max_j L_j$$

Algorithm configuration Our algorithm is detailed in Algorithm 1. In this experiment, we compare our approach against the accelerated proximal method, the optimistic method, and the extragradient method, defined as follows:

Accelerated Proximal Gradient Descent When $\gamma > 0$, the exact maximization over λ yields a strongly-smoothed primal function $g(w) = \max_{\lambda \in \Delta_J} \Phi(w, \lambda)$. Compute

$$\lambda_t^\star = \Pi_{\Delta_J}\left(\mathbf{1}_J + \frac{f(y_{t-1})}{\gamma}\right)$$
$$L_{\text{local}} = L + \frac{\|\nabla f(y_{t-1})\|_2^2}{\gamma}$$
$$w_t = y_{t-1} - \frac{1}{L_{\text{local}}}\nabla f(y_{t-1})\lambda_t^\star$$

Apply Nesterov Momentum Update:

$$\tau_t = \frac{1 + \sqrt{1 + 4\tau_{t-1}^2}}{2}$$
$$y_t = w_t + \frac{\tau_{t-1} - 1}{\tau_t}(w_t - w_{t-1})$$

Extragradient Compute the step-size and midpoint

$$\eta_t = \frac{1}{\max(L, \gamma) + \|\nabla f(w_{t-1})\|_2}$$
$$w_{\text{mid}} = w_{t-1} - \eta_t \nabla f(w_{t-1})\lambda_{t-1}$$
$$\lambda_{\text{mid}} = \Pi_{\Delta_J}\left[\lambda_{t-1} + \eta_t \nabla_\lambda \Phi(w_{t-1}, \lambda_{t-1})\right]$$

Compute the final update

$$w_t = w_{t-1} - \eta_t \nabla f(w_{\text{mid}}), \lambda_{\text{mid}}$$
$$\lambda_t = \Pi_{\Delta_J}\left[\lambda_{t-1} + \eta_t \nabla_\lambda \Phi(w_{\text{mid}}, \lambda_{\text{mid}})\right]$$