From the simplex to the sphere: Faster constrained optimization using the Hadamard parametrization

Qiuwei Li[*] Daniel McKenzie[†] Wotao Yin[*]

December 13, 2021

Abstract

We show how to convert the problem of minimizing a convex function over the standard probability simplex to that of minimizing a nonconvex function over the unit sphere. We prove the landscape of this nonconvex problem is benign, *i.e.* every stationary point is either a strict saddle or a global minimizer. We exploit the Riemannian manifold structure of the sphere to propose several new algorithms for this problem. When used in conjunction with line search, our methods achieve a linear rate of convergence for non-degenerate interior points, both in theory and in practice. Extensive numerical experiments compare the performance of our proposed methods to existing methods, highlighting the strengths and weaknesses. We conclude with recommendations for practitioners.[1]

1 Introduction

In this paper we are primarily interested in the problem:

$$x^\star \in \operatorname*{argmin}_{x \in \Delta_n} f(x) \tag{C_1}$$

where

$$\Delta_n = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1 \text{ and } x_i \geq 0, \ \forall i \right\} \tag{1}$$

is the *probability simplex* and f is convex and twice continuously differentiable (henceforth: *smooth*). We also consider several other related geometric constraints and the more general case where f is nonconvex (See Section 7). Optimization over Δ_n is a crucial step of many modern applications in machine learning, signal processing, control and game theory. Typical applications include, but are not limited to, estimation of mixture proportions (Keshava, 2003), probability density estimation (Bunea et al., 2010), convex aggregation learning (Nemirovski, 2000), training of support vector machines (Clarkson, 2010), portfolio optimization (Bomze, 2002), population dynamics (Zeeman, 1980), graph theory (De Klerk, 2008), Bayesian neural architecture (Limmer and Stańczak, 2018), and archetypal analysis (Bauckhage et al., 2015). A natural approach to this problem is *Projected Gradient Descent (PGD)*:

$$x^{(k+1)} = P_{\Delta_n}\left(x^{(k)} - \eta_k \nabla f(x^{(k)}) \right) \tag{2}$$

where η_k is an appropriately chosen step size and $P_{\Delta_n}(y)$ is the projection onto Δ_n,

$$P_{\Delta_n}(y) := \operatorname*{argmin}_{x \in \Delta_n} \|x - y\|^2 \tag{3}$$

[*] Alibaba DAMO Academy, Bellevue, WA
[†] University of California, Los Angeles, CA

[1] All code available at https://github.com/DanielMckenzie/HadRGD

Standard theory (*e.g.* (Bertsekas, 1997, Chpt. 3)) shows that when f is convex (resp. strong convex) PGD finds an iterate satisfying $f(x_k) - f^\star \leq \varepsilon$ (henceforth: an ε-optimal solution) in $\mathcal{O}(1/\varepsilon)$ (resp. $\mathcal{O}(\log(1/\varepsilon))$) iterations. In many cases (*e.g.* simplex-constrained least-squares, see Section 8) computing P_Δ is a substantial part of the per-iteration computational cost. Theoretically, $P_{\Delta_n}(x)$ can be computed with worst case complexity $\mathcal{O}(n)$. However, this involves either (i) using a linear time median selection algorithm Blum et al. (1973) known to be slow in practice; or (ii) using the recently proposed algorithm of Perez et al. (2020), which requires *a priori* knowledge (*i.e.* a bound on the size of the entries of x) as well as some non-standard floating point operations. Hence, randomized algorithms (Duchi et al., 2008; Condat, 2016) with worst case complexity $\mathcal{O}(n^2)$ but empirically observed complexity closer to $\mathcal{O}(n)$ are frequently used in practice. To summarize, the total computational complexity of finding an ε-optimal solution to (C_1) is between $\mathcal{O}(n/\varepsilon)$ and $\mathcal{O}(n^2/\varepsilon)$ ($\mathcal{O}(n\log(1/\varepsilon))$ and $\mathcal{O}(n^2\log(1/\varepsilon))$ if f is strongly convex) but most implementations of P_{Δ_n} used in practice yield a worst-case complexity of $\mathcal{O}(n^2/\varepsilon)$ (again, $\mathcal{O}(n^2\log(1/\varepsilon))$ if f is strongly convex). We note there are are many algorithms for (C_1) besides PGD (see Section 2) but they all have similar computational complexities.

In this work we propose a new approach to (C_1). When f satisfies a non-degeneracy condition slightly weaker than strong convexity, our approach has worst-case complexity $\mathcal{O}(n\log(1/\varepsilon))$ *and* is fast in practice. Our approach also avoids the sorting or median selection sub-routines required by many algorithms for P_{Δ_n}. As these are tricky to parallelize, our approach may prove to be more GPU-friendly.

Hadamard parametrization Our proposed approach is conceptually simple. We use the *Hadamard parametrization*: $x = z \odot z$, where $\odot$ represents the elementwise product $(u \odot v)_k = u_k v_k$. Note

$$\sum_i x_i = 1 \text{ and } x_i \geq 0,\ \forall i \iff x = z \odot z, \|z\|_2^2 = 1$$

As a consequence, the probability simplex constraint $x \in \Delta_n$ in the original space is transformed to a unit sphere constraint $z \in \mathcal{S}_n$ in the Hadamard-parametrization space, where $\mathcal{S}_n := \{z \in \mathbb{R}^n : \|z\|_2 = 1\}$ denotes the unit sphere. (See Fig. 1).

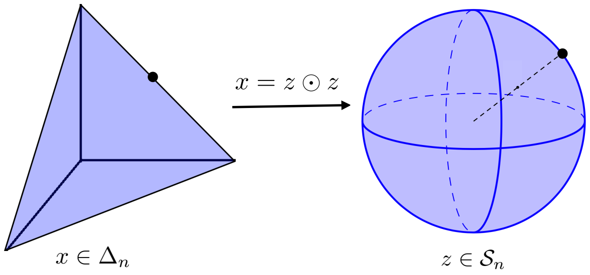

Figure 1: $x \in \Delta_n$ in the original space is transformed as $z \in \mathcal{S}_n$ in the Hadamard-parametrization space.

Using the Hadamard parametrization, we transform (C_1) into the following nonconvex problem

$$z^\star \in \operatorname*{argmin}_{z \in \mathcal{S}_n} \{g(z) := f(z \odot z)\} \tag{NC$_1$}$$

Importantly $\mathcal{S}_n$ is a Riemannian manifold, unlike Δ_n. So, techniques from Riemannian optimization (Absil et al., 2009) may be applied to (NC_1). The function g is nonconvex, hence it is not *a priori* obvious if (Riemannian) gradient based methods applied to (NC_1) will find a global minimizer, or at what rate. We build on recent research on nonconvex optimization to show:

Theorem (Theorem 1, informally stated)**.** *Every stationary point of $g(z)$ is either a strict saddle or a global minimizer.*

It is easily checked that if $z^\star$ is a global minimizer of $g(z)$ then $x^\star = z^\star \odot z^\star$ is a global minimizer of $f(x)$. Thus, *any* method applied to Problem (NC$_1$) which converges to a second order stationary point provably solves Problem (C$_1$). We analyze several such methods, namely Perturbed Riemannian Gradient Descent (PRGD) (Criscitiello and Boumal, 2019) and two variations of RGD with line search. Defining $\mathcal{U} := \{u \in \mathbb{R}^n : \sum_i u_i = 0\}$ we show

Theorem (Theorems 4 and 6, informally stated)**.** *Suppose $x^\star$ is in the interior of Δ_n and $\nabla^2 f(x^\star)$ is positive definite (henceforth: PD) when restricted to $\mathcal{U}$. Then:*

1. *PRGD finds an ε-optimal solution in $\mathcal{O}(\log^4(n)/\varepsilon)$ iterations and $\mathcal{O}(n\log^4(n)/\varepsilon)$ total operations.*
2. *RGD with a backtracking, Armijo-Wolfe linesearch finds an ε-optimal solution in $\mathcal{O}(\log(1/\varepsilon))$ iterations and $\mathcal{O}(n\log(1/\varepsilon)$ total operations.*

Our experimental results show that, when the step-size is well tuned, PRGD is significantly faster than PGD. Finally, we show experimentally that RGD with a non-monotone line search incorporating the Barzilai-Borwein step-size rule is the overall fastest algorithm for Problem (C$_1$) that we tested, outperforming PGD and other algorithms by an order of magnitude in some cases.

Notation We use x to denote a variable constrained to Δ_n, while z denotes a variable in the Hadamard space. Our original objective function will always be f and by g we always mean the Hadamard parametrized version of f: $g(z) = f(z \odot z)$. The notations ∇ and grad will be reserved for (Euclidean) gradient operator and Riemannian gradient operator, respectively. We will use ∇^2 and Hess to denote the (Euclidean) Hessian and Riemannian Hessian operator, respectively. Finally, by $\mathrm{int}(\Delta_n)$ we mean the interior of Δ_n.

Hadamard Calculus For the reader's convenience, we recall a few properties of the Hadamard product. We defer all proofs to the appendix.

(H1) $1_n \odot z = z$ where 1_n is the all-one vector in $\mathbb{R}^n$

(H2) If $d \odot z \odot z = 0$ then $d \odot z = 0$

(H3) $\mathrm{diag}(z)d = z \odot d$ and $d^\top \mathrm{diag}(z)d = \langle d, z \odot d\rangle = \langle z, d \odot d\rangle$

(H4) $\|d \odot z\|_2 \leq \|d\|_2 \|z\|_\infty$

(H5) $\nabla g(z) = 2\nabla f(x) \odot z$ and $\nabla^2 g(z) = 2\mathrm{diag}(\nabla f(x)) + 4\mathrm{diag}(z)\nabla^2 f(x)\mathrm{diag}(z)$

2 Prior Work

Hadamard parametrization There has been a flurry of recent papers examining the Hadamard parametrization. In particular, we highlight the papers (Vaskevicius et al., 2019; Zhao et al., 2019) which study the constrained convex problem:

$$\underset{x}{\text{minimize}}\ \|x\|_1 \text{ subject to } Ax = b, \tag{4}$$

frequently used to find the sparsest solution to an underdetermined system $Ax = b$. Informally, their approach is to set $x = z \odot z$ and then apply carefully initialized gradient descent to the *unconstrained* nonconvex problem:

$$\underset{z}{\text{minimize}}\ \|Az \odot z - b\|_2^2 \tag{5}$$

Using results on the implicit bias of gradient descent (Ali et al., 2019; Bauer et al., 2007; Bühlmann and Yu, 2003), they argue gradient descent applied to (5) will find a $z^\star \in \{z : Az \odot z = b\}$ of minimal ℓ_2 norm. As $\|z\|_2^2 = \|z \odot z\|_1$ they show $x^\star := z^\star \odot z^\star$ is indeed the solution to (4). At the risk of over-generalizing, this line of research can be summarized as using the Hadamard parametrization to exploit implicit regularization.

Although inspired by these works, our approach is distinct from this line of research. Instead of using the Hadamard parametrization to take advantage of implicit regularization, we use it to convert a non-smooth constraint set into a smooth one.

Simplex Minimization Alternative approaches to solving (C_1) include the Frank-Wolfe algorithm (Frank et al., 1956; Jaggi, 2013), also known as the conditional gradient algorithm, and mirror descent (Ben-Tal et al., 2001; Beck and Teboulle, 2003), known in this context as the entropic mirror descent algorithm or the exponentiated gradient algorithm. When f is strongly convex certain variants of Frank-Wolfe achieve a linear convergence rate (Lacoste-Julien and Jaggi, 2015; Pedregosa et al., 2020) although for merely convex f the convergence rate is sublinear. Interior point methods (*e.g.* (Koh et al., 2007)) enjoy a fast convergence rate but, as the cost of each iteration is typically quadratic in n, are intractible for high-dimensional problems.

3 Riemannian Optimization

We recall a few notions required for discussing optimization on Riemannian manifolds, specialized to the case of the sphere, $\mathcal{S}_n$. For any $z \in \mathcal{S}_n$ the *tangent space* is $T_z\mathcal{S}_n = \{v \in \mathbb{R}^n : \ v^\top z = 0\}$. Let Proj_z denote the projection onto $T_z\mathcal{S}_n$. One can verify that $\mathrm{Proj}_z(w) = w - (w^\top z)z$. For any smooth $f : \mathcal{S}_n \to \mathbb{R}$ the *Riemannian gradient* at $z \in \mathcal{S}_n$ is the projection of the regular gradient onto $T_z\mathcal{S}_n$: $\mathrm{grad}_z\, f = \mathrm{Proj}_z \nabla f(z)$. Similarly, we may define the *Riemannian Hessian* at $z \in \mathcal{S}_n$ as the operator $\mathrm{Hess}\, f(z) = \mathrm{Proj}_z \circ \left(\nabla^2 f(z) - \nabla f(z)^\top z\right) \circ \mathrm{Proj}_z$; see (Boumal, 2020, Sec. 7) for further details.

Given $z \in \mathcal{S}_n$ and $v \in T_z\mathcal{S}_n$ with $\|v\| = 1$ there exists a unique *geodesic* emanating from z in the direction of v, which is a curve on $\mathcal{S}_n$ written as $\gamma_{z,v}(t) : \mathbb{R} \to \mathcal{S}_n$. These generalize the role of straight lines in Euclidean geometry. On $\mathcal{S}_n$ the geodesics are precisely the great circles. It is convenient to define the *exponential map* at $z \in \mathcal{S}_n$:

$$\begin{aligned} \exp_z : & T_z\mathcal{S}_n \to \mathcal{S}_n \\ & v \mapsto \gamma_{z,\hat{v}}(\|v\|) \text{ where } \hat{v} := v/\|v\| \end{aligned}$$

(RGD) mimics regular (*i.e.* Euclidean) gradient descent, except instead of using $\nabla f(x_k)$ we use $\mathrm{grad}\, f(x_k)$, and instead of stepping along the straight line in the direction $-\nabla f(x_k)$ we move along the geodesic in the direction $-\,\mathrm{grad}\, f(x_k)$:

$$x_{k+1} = \exp_{x_k}(-\alpha_k\, \mathrm{grad}\, f(x_k)) \tag{6}$$

By construction $x_{k+1} \in \mathcal{S}_n$ hence RGD is a *feasible algorithm*. Algorithm 1, which we dub HadRGD, describes how to apply RGD to (C_1) via the Hadamard parametrization. Note the function which RGD is actually minimizing, g, will in general be nonconvex (even if f is convex). However g does inherit the smoothness of f:

Lemma 3.1. *Suppose f is L-Lipschitz differentiable. Then g is $\tilde{L}$-Lipschitz differentiable with $\tilde{L} = 4L + 2M$ where $M = \max_{x\in\Delta_n} \|\nabla f(x)\|_\infty$.*

As $\nabla f(x)$ is continuous and Δ_n is compact, $M < \infty$.

4 Landscape Analysis

Recall in this paper, we transform the simplex constrained convex optimization problem (C_1) to the unit ball constrained nonconvex optimization problem (NC_1). Therefore, the Karush–Kuhn–Tucker (KKT) conditions can be used to characterize the (global) optimality conditions of (C_1). In this section we establish the one-to-one correspondence between the global minimizers of (C_1) and the second-order KKT points of the Hadamard parameterized nonconvex optimization problem (NC_1).

KKT conditions of (C$_1$) Recall the Lagrangian is

$$\mathcal{L}_C(x, \lambda, \beta) = f(x) - \lambda(1_n^\top x - 1) + \beta^\top x \tag{7}$$

Since (C$_1$) is convex, the global optimality conditions are given by the (first-order) KKT conditions: there exist $\lambda^\star \in \mathbb{R}$ and $\beta^\star \in \mathbb{R}^n$ such that

$$\begin{aligned} \nabla f(x^\star) &= \lambda^\star 1_n + \beta^\star && \text{(8a)} \\ x^\star &\geq 0 && \text{(8b)} \\ \beta^\star &\geq 0 && \text{(8c)} \\ x^\star \odot \beta^\star &= 0 && \text{(8d)} \\ 1_n^\top x^\star &= 1 && \text{(8e)} \end{aligned}$$

First-order KKT conditions of (NC$_1$) The Lagrangian of (NC$_1$) is

$$\mathcal{L}_N(z, \lambda_N) = g(z) - \lambda_N(\|z\|_2^2 - 1) \tag{9}$$

The first-order optimality conditions of (NC$_1$) are given by the (first-order) KKT conditions: there exists $\lambda_N^\star \in \mathbb{R}$ such that

$$\begin{aligned} \nabla f(z^\star \odot z^\star) \odot z^\star &= \lambda_N^\star z^\star && \text{(10a)} \\ \|z^\star\|_2^2 &= 1 && \text{(10b)} \end{aligned}$$

A point $z^\star$ satisfying the first-order optimality conditions is called a KKT/stationary point.

Second-order KKT conditions of (NC$_1$) Using (H5), the second-order optimality conditions of (NC$_1$) are given by the second-order KKT conditions: there exists $\lambda_N^\star \in \mathbb{R}$ such that $(z^\star, \lambda_N^\star)$ satisfies (10a) and (10b) and for all $d \perp z^\star$,

$$\begin{aligned} & d^\top \left[\nabla_z^2 \mathcal{L}_N(z^\star, \lambda_N^\star)\right] d \geq 0 \\ \Longleftrightarrow\ & d^\top \Big[2\mathrm{diag}(\nabla f(z^\star \odot z^\star)) \\ & \quad + 4\mathrm{diag}(z^\star)\nabla^2 f(z^\star \odot z^\star)\mathrm{diag}(z^\star) - 2\lambda^\star I_n \Big] d \geq 0 \\ \overset{\text{(H3)}}{\Longleftrightarrow}\ & 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda^\star \|d\|^2 \\ & + 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) \geq 0 \end{aligned} \tag{11}$$

where $\nabla_z^2(\cdot)$ denotes the partial Hessian with respect to z. A point $z^\star$ satisfying the second-order optimality conditions is a second-order KKT/stationary point.

Remark 1. *Note that $z^\star$ is a second order stationary point in the Riemannian sense,* i.e. $\mathrm{grad}(f)(z^\star) = 0$ *and* $d^\top \,\mathrm{Hess}\, f(z^\star) d \geq 0$ *for all $d \in T_z^\star \mathcal{S}_n$, if and only if $z^\star$ is a second order KKT point (Luenberger, 1972; Boumal, 2020). This justifies using these terms interchangeably.*

Strict saddle points of (NC$_1$) Suppose $z^\star$ is a stationary point violating the second-order optimality conditions, *i.e.* there exists $d \perp z^\star$ such that

$$d^\top \left[\nabla_z^2 \mathcal{L}_N(z^\star, \lambda_N^\star)\right] d < 0. \tag{12}$$

We call such a stationary point $z^\star$ a *strict saddle point.* Note that any stationary point must be either a strict saddle or a second-order stationary point. Remarkably, many existing iterative algorithms are able to avoid strict saddle points and converge to second-order stationary points, *e.g.* Trust-region method (Sun et al., 2015), cubic-regularization method (Zhang and Zhang, 2018; Nesterov and Polyak, 2006), Riemannian gradient descent (Criscitiello and Boumal, 2019), and projected gradient descent (Ge et al., 2015).

We now provide our main landscape analysis results.

Theorem 1. *Suppose f is convex and $z^\star$ is any stationary point $z^\star$ of* (NC$_1$)*. Then $z^\star$ is a second-order stationary point of* (NC$_1$) *if and only if $z^\star \odot z^\star$ is a global minimizer of* (C$_1$)*.*

Proof. Define the Hadamard decomposition set as $\mathcal{Z}^\star = \{z^\star : z^\star \odot z^\star = x^\star\}$, where $x^\star$ is any KKT point of (C$_1$). Then Theorem 1 is equivalent to saying that $z^\star$ is a second order stationary point of (NC$_1$) if and only if $z^\star \in \mathcal{Z}^\star$.

1. **"If" part:** If $z^\star \in \mathcal{Z}^\star$, then $z^\star$ is a second order stationary point of (NC$_1$). Suppose $z^\star \in \mathcal{Z}^\star$. We first show $z^\star$ satisfies the first-order optimality conditions (10a) and (10b). Multiplying both sides of the optimality condition (8a) by $z^\star$:

$$\nabla f(x^\star) \odot z^\star = \lambda^\star 1_n \odot z^\star + \beta^\star \odot z^\star \tag{13}$$

$$\overset{\text{(H1)}}{\Longrightarrow} \nabla f(z^\star \odot z^\star) \odot z^\star = \lambda^\star z^\star + \beta^\star \odot z^\star \tag{14}$$

By complementary slackness (8d): $z^\star \odot z^\star \odot \beta^\star = 0$, hence $z^\star \odot \beta^\star = 0$ by (H2). Thus (14) reduces to (10a) by choosing $\lambda_N^\star = \lambda^\star$. Note (8e) is equivalent to (10b) as

$$1 = 1_n^\top x^\star = 1_n^\top z^\star \odot z^\star = \sum_{i=1}^n (z_i^\star)^2 = \|z^\star\|_2^2.$$

It remains to show the second-order optimality conditions (11) of (NC$_1$) with $\lambda_N^\star = \lambda^\star$. Since $x^\star$ satisfies (8a) and (8c), we have

$$\begin{aligned} & \nabla f(x^\star) = \lambda^\star 1_n + \beta^\star, \beta^\star \geq 0 \\ \Longleftrightarrow\ & \operatorname{diag}(\nabla f(z^\star \odot z^\star)) - \lambda^\star I_n = \operatorname{diag}(\beta^\star) \succeq 0 \\ \Longleftrightarrow\ & d^\top \left[\operatorname{diag}(\nabla f(z^\star \odot z^\star)) - \lambda^\star I_n\right] d \geq 0,\ \forall d \\ \overset{\text{(H3)}}{\Longleftrightarrow}\ & \langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - \lambda^\star \|d\|^2 \geq 0,\ \forall d \end{aligned}$$

Therefore, we obtain that

$$\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - \lambda^\star \|d\|^2 \geq 0,\ \forall d \tag{15}$$

Plugging this into (11), we get for any $d \in \mathbb{R}^n$

$$\begin{aligned} & d^\top [\nabla_z^2 \mathcal{L}_N(z^\star, \lambda^\star)] d \\ =& 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) \\ & + 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda^\star \|d\|^2 \\ \geq& 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) \geq 0 \end{aligned} \tag{16}$$

which follows from the convexity assumption of f.

2. **"Only if" part:** $z^\star$ is second order stationary points of (NC$_1$) only if $z^\star \in \mathcal{Z}^\star$. For convenience, we show its contrapositive: If $z^\star \notin \mathcal{Z}^\star$, then $z^\star$ is a strict saddle point of (NC$_1$). Because $z^\star$ is a stationary point of (NC$_1$), it satisfies (10b):

$$z^\star \odot z^\star \geq 0 \text{ and } 1_n^\top (z^\star \odot z^\star) = 1$$

That is, $z^\star \odot z^\star$ satisfies the optimality conditions (8b) and (8e) of (C_1). Since $z^\star$ satisfies (10a), there exists $\lambda_N^\star \in \mathbb{R}$ such that

$$\begin{aligned} \nabla f(z^\star \odot z^\star) \odot z^\star &= \lambda_N^\star z^\star \\ \Longrightarrow [\nabla f(z^\star \odot z^\star)]_k &= \lambda_N^\star, \ \forall z_k^\star \neq 0 \end{aligned}$$

On the other hand, (8a),(8c) and (8d) are equivalent to that there exists some $\lambda^\star$ such that

$$\begin{cases} [\nabla f(x^\star)]_k = \lambda^\star, \ \forall x_k^\star \neq 0 \\ [\nabla f(x^\star)]_k \geq \lambda^\star, \ \forall x_k^\star = 0 \end{cases}$$

For the sake of contradiction, further suppose

$$[\nabla f(z^\star \odot z^\star)]_k \geq \lambda_N^\star, \ \forall z_k^\star = 0$$

then $z^\star \odot z^\star$ satisfies the optimality conditions (8a),(8c) and (8d) by choosing $\lambda^\star = \lambda_N^\star$ and $\beta_k^\star = [\nabla f(z^\star \odot z^\star)]_k - \lambda_N^\star \geq 0$. Consequently, $z^\star \odot z^\star$ satisfies the entire KKT conditions (8a)–(8e), and so $z^\star \in \mathcal{Z}^\star$, which is a contradiction to the assumption $z^\star \notin \mathcal{Z}^\star$. Therefore,

$$[\nabla f(z^\star \odot z^\star)]_{k^\star} < \lambda_N^\star \quad \text{for some } z_{k^\star}^\star = 0 \tag{17}$$

By constructing $d = e_{k^\star}$, where $\{e_i\}$ denotes the canonical basis in $\mathbb{R}^n$, we have $\langle d, z^\star\rangle = 0$ and $d \odot z^\star = 0$. By direct computation and (11) we obtain

$$\begin{aligned} & d^\top [\nabla^2 \mathcal{L}_N(z^\star, \lambda_N^\star)] d \\ =& 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) \\ & + 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda_N^\star \|d\|^2 \\ =& 0 + 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda_N^\star \|d\|^2 \\ =& 2[\nabla f(z^\star \odot z^\star)]_{k^\star} - 2\lambda_N^\star \\ \overset{(17)}{<}& 0 \end{aligned}$$

Therefore, $z^\star$ is a strict saddle point of (NC_1).

Non-degenerate stationary points It will also be useful to consider the following refined version of a second-order stationary point.

Definition 1 (Definition 3.1 (Yang, 2007)). *$z^\star \in \mathcal{S}_n$ is a nondegenerate stationary point of g if the first-order optimality conditions (10a) and (10b) are satisfied, and the second-order optimality condition (11) holds with strict inequality.*

Roughly speaking, a nondegenerate stationary point is an isolated local minimizer. We now characterize nondegenerate stationary points of $g(z) := f(z \odot z)$.

Theorem 2. *Suppose $x^\star$ is a global minimizer of (C_1) with $x^\star \in \operatorname{int}(\Delta_n)$ and $\nabla^2 f(x^\star)$ PD when restricted to $\mathcal{U} := \{u \in \mathbb{R}^n : \sum_i u_i = 0\}$. Then all the coresponding Hadamard parametrizations $z^\star \in \mathcal{Z}^\star$ are nondegenerate stationary points of (NC_1).*

5 Perturbed Riemannian Gradient Descent

As we are using RGD to minimize a *nonconvex* function, care must be taken when analyzing its convergence. First recall:

Algorithm 1 HadRGD for (C$_1$)

```
Input: $x_0 \in \Delta_n$: initial point, $\alpha$: step size, $K$: number of iterations, $f$: original objective function
$z_0 = \sqrt{x_0}$ {(Defined componentwise)}
$g(z) := f(z \odot z)$
for k=1,..., K do
    $z_{k+1} = \exp_{x_k}(-\alpha \operatorname{grad} g(z_k))$
end for
Return $x_K = z_K \odot z_K$
```

Definition 2 ((Criscitiello and Boumal, 2019))**.** *A point $z \in \mathcal{S}_n$ is an ϵ-second-order stationary point of the twice-differentiable function $g : \mathcal{S}_n \to \mathbb{R}$ if*

$$\| \operatorname{grad} g(z) \| \leq \epsilon \quad \textit{and} \quad \lambda_{\min}(\operatorname{Hess} g(z)) \geq -\sqrt{\rho\epsilon}$$

where $\lambda_{\min}(H)$ denotes the smallest eigenvalue of the symmetric operator H and ρ denotes the Lipschitz constant of the Hessian of the pullback of g from the manifold to tangent space.

There is no guarantee RGD applied to a nonconvex function will converge to a second order stationary point (it may find a saddle point). Fortunately, (Criscitiello and Boumal, 2019) shows a *Perturbed* version of RGD (PRGD, c.f. Algorithm 1 in (Criscitiello and Boumal, 2019)) will, with high probability (w.h.p), find an ϵ-second-order stationary point. We reproduce their result, adapted to the sphere, here.

Theorem 3. *PRGD applied to $g : \mathcal{S}_n \to \mathbb{R}$ finds an ϵ-second-order stationary point of $g(z)$ w.h.p. in $\mathcal{O}\left((\log n)^4/\epsilon^2\right)$ iterations.*

We call the combination of the Hadamard parametrization and PRGD *HadPRGD* (see Algorithm 2). As a consequence of our landscape analysis (Theorems 1 and 2):

Theorem 4. *Suppose $x^\star \in \operatorname{int}(\Delta_n)$ and $\nabla^2 f(x^\star)$ PD when restricted to $\mathcal{U} := \{u \in \mathbb{R}^n : \sum_i u_i = 0\}$ for every global minimizer of (C$_1$). Then HadPRGD finds an ε-optimal solution to (C$_1$) w.h.p. in $\mathcal{O}((\log n)^4/\varepsilon)$ iterations, for ε small enough.*

Each iteration of HadPRGD (excluding the perturbed steps, which are hardly ever triggered) requires $\mathcal{O}(n)$ operations, hence the total flop count of HadPRGD is $\mathcal{O}\left(n(\log n)^4/\varepsilon\right)$. Although this is not an improvement on the complexity of PGD for (C$_1$), we find experimentally that HadPRGD tolerates much larger step-sizes, leading to faster convergence in practice. To substantiate this we consider the simplex-constrained least squares problem

$$x^\star = \underset{x \in \Delta_n}{\operatorname{argmin}} \left\{ f(x) = \|Ax - b\|_2^2 \right\} \tag{18}$$

where $A \in \mathbb{R}^{m \times n}$ with $m = 0.1n$ and $b = Ax_{\text{true}}$ for randomly selected $x_{\text{true}} \in \operatorname{int}(\Delta_n)$. We record, as a function of n, the number of iterations and wall-clock time required by HadPRGD and PGD to reach to find an ε-optimal solution with $\varepsilon = 10^{-16}$ (see Figure 2). The step sizes of HadPRGD and PGD are hand-tuned to be as large as possible while still converging.

6 RGD with Line Search

Clearly, HadRGD is effective when using an appropriately large step-size, but it is unclear from theoretical grounds how large this step-size can be. Thus, we implement HadRGD in conjunction with a line search algorithm. We consider two approaches.

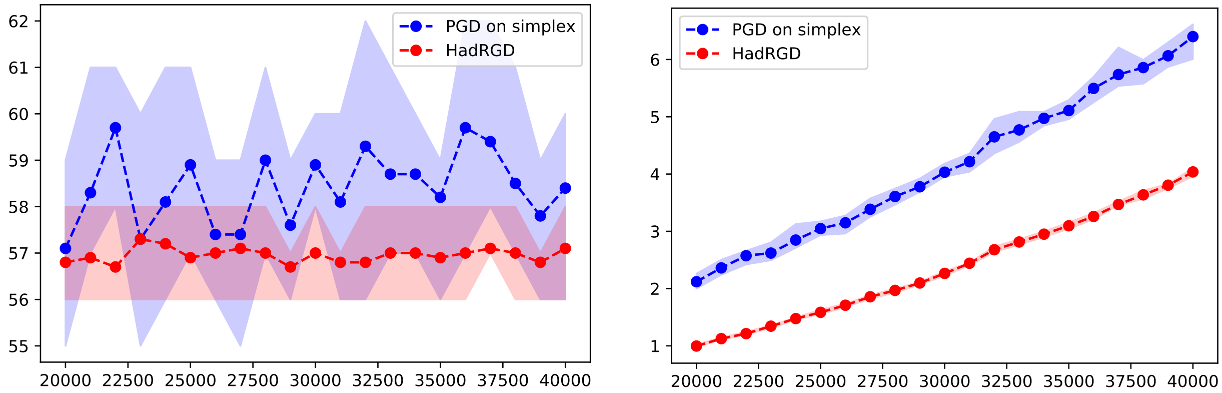

Figure 2: **Left:** Number of iterations vs. n. **Right:** Wall-clock time vs. n. When the step-size is properly tuned, PRGD requires fewer iterations than PGD on the simplex. Moreover, the per-iteration cost of PRGD is lower, leading to a lower overall run-time.

Armijo-Wolfe Suppose $v \in T_z\mathcal{S}_n$ is a *descent direction*[2] for g. We say $\alpha^\star$ satisfies the (Riemannian) Armijo-Wolfe conditions along the geodesic traced out by $\exp_z(\alpha v)$ if

$$g\left(\exp_z(\alpha^\star v)\right) \leq g\left(z\right) + \rho_1 \alpha^\star \langle \operatorname{grad} g(z), v\rangle \tag{19a}$$

$$g'\left(\exp_z(\alpha^\star v)\right) \geq \rho_2 \langle \operatorname{grad} g(z), v\rangle \tag{19b}$$

where $g'\left(\exp_z(\alpha v)\right)$ is shorthand for the derivative of $\alpha \mapsto g\left(\exp_z(\alpha v)\right)$. Informally, (19a) guarantees sufficient descent while (19b) guarantees approximate stationarity (recalling $\langle \operatorname{grad} g(z), v\rangle < 0$). We implement HadRGD, with α_k chosen to satisfy the Armijo-Wolfe conditions along $\exp_{z_k}(\alpha \operatorname{grad} g(z_k))$ via backtracking line search, as HadRGD-AW (Alg. 4 in Appendix E).

Barzilai-Borwein (BB) The BB step-size[3] rule is

$$\alpha_k^{\mathrm{BB}} = \|s_{k-1}\|_2^2 / \langle s_{k-1}, y_{k-1}\rangle \tag{20}$$

where $s_{k-1} = z_k - z_{k-1}$ and $y_{k-1} = \operatorname{grad} g(z_k) - \operatorname{grad} g(z_{k-1})$. HadRGD with the BB step-size need not be monotone, thus we follow (Wen and Yin, 2013; Zhang and Hager, 2004) and determine α_k via a non-monotone line search starting at α_k^{BB}. That is, we select $z_{k+1} = \exp_{z_k}(\alpha_k \operatorname{grad} g(z_k))$ where $\alpha_k = \delta^h \alpha_k^{\mathrm{BB}}$ where h is the smallest integer satisfying

$$g(\exp_{z_k}(\alpha_k \operatorname{grad} g(z_k))) \leq C_k - \rho_1 \alpha_k \|\operatorname{grad} g(z_k)\|^2$$

where C_k is a running average:

$$C_{k+1} = \frac{\eta Q_k C_k + g(z_{k+1})}{Q_{k+1}} \quad \text{and} \quad Q_{k+1} = \eta Q_k + 1$$

We implement this as HadRGD-BB (Alg. 5 in Appendix E).

All parameters $(\rho_1, \rho_2, \delta, \eta, \ldots)$ are set to the values suggested in (Wen and Yin, 2013). See Appendix E for further implementation details.

[2] *i.e.* $\langle \operatorname{grad} g(z), v\rangle < 0$

[3] There is another, closely related BB step-size rule. See (Wen and Yin, 2013) for discussion

6.1 Linear Convergence for Interior Points

We show HadRGD-AW converges at a linear rate.

Theorem 5 (Theorem 4.1 (Yang, 2007), adapted)**.** *Suppose $z^\star \in \mathcal{S}_n$ is a non-degenerate stationary point of g. Let $\{z^k\}$ be a sequence of points converging to $z^\star$ constructed by HadRGD-AW. Then, there exists a constant E' such that, for some integer $K_0 \geq 0$ and $\theta \in (0,1)$, we have $\|z^{k+K_0} - z_\star\|_2 \leq E'\theta^{k/2}$.*

Combining Theorems 5 and 2:

Theorem 6. *Suppose all global minimizers $x^\star$ of (C$_1$) are in $\mathrm{int}(\Delta_n)$ and have $\nabla^2 f(x^\star)$ positive definite (PD). Then HadRGD-AW finds an ε-optimal solution in $\mathcal{O}(\log(1/\varepsilon))$ iterations.*

Each iteration of HadRGD-AW requires $\mathcal{O}(n)$ operations yielding a total flop count of $\mathcal{O}(n\log(1/\varepsilon)$. We know of no corresponding proof for HadRGD-BB, but empirically HadRGD-BB is significantly faster.

7 Extensions

7.1 Extension to other related geometries

We extend our framework from Δ_n to several related geometries, deferring all proofs to Appendix F.

The unit simplex Let $\blacktriangle_n$ denote the unit simplex

$$\blacktriangle_n := \left\{ x \in \mathbb{R}^n : \sum_i x_i \leq 1 \text{ and } x_i \geq 0, \ \forall i \right\} \tag{21}$$

Note that the boundary of the unit simplex $\blacktriangle_n$ is the probability simplex Δ_n. We now consider

$$x^\star \in \operatorname*{argmin}_{x \in \blacktriangle_n} f(x) \tag{C$_2$}$$

Using $x = z \odot z$ we transform (C$_2$) to a nonconvex problem, this time over the unit ball $\mathcal{B}_n^2 := \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$:

$$z^\star \in \operatorname*{argmin}_{z \in \mathcal{B}_n^2} f(z \odot z) \tag{NC$_2$}$$

Similar to the Δ_n case, (NC$_2$) has a benign landscape.

Theorem 7. *Suppose f is convex and $z^\star$ is any stationary point $z^\star$ of (NC$_2$). Then $z^\star$ is a second-order stationary point of (NC$_2$) if and only if $z^\star \odot z^\star$ is a global minimizer of (C$_2$).*

The weighted probability simplex For any $a \in \mathbb{R}^n$ with $a > 0$ (entrywise), consider the weighted probability simplex:

$$\Delta_n^a := \left\{ z \in \mathbb{R}^n : \sum_{i=1}^n a_i z_i = 1 \text{ and } z_i \geq 0, \ \forall i \right\}. \tag{22}$$

Let us consider the following optimization problem:

$$x^\star \in \operatorname*{argmin}_{x \in \Delta_n^a} f(x) \tag{C$_3$}$$

Using once more $x = z \odot z$, we obtain:

$$z^\star \in \operatorname*{argmin}_{z \in \mathcal{B}_n^a} f(z \odot z) \tag{NC$_3$}$$

Here, the unit a-weighted ℓ_2 norm ball is defined as

$$\mathcal{B}_n^a := \left\{x \in \mathbb{R}^n : \|x\|_{\mathrm{diag}(a)}^2 = 1\right\},$$

where

$$\|x\|_{\mathrm{diag}(a)}^2 := \sum_i a_i x_i^2.$$

We claim (NC$_3$) enjoys a benign landscape.

Theorem 8. *Suppose f is convex and $z^\star$ is any stationary point $z^\star$ of* (NC$_3$)*. Then $z^\star$ is a second-order stationary point of* (NC$_3$) *if and only if $z^\star \odot z^\star$ is a global minimizer of* (C$_3$)*.*

The ℓ_1 norm ball Consider the minimization:

$$x^\star \in \operatorname*{argmin}_{x \in \mathcal{B}_n^1} f(x) \tag{C$_4$}$$

where the unit ℓ_1 ball is defined as

$$\mathcal{B}_n^1 := \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}.$$

Since x may have negative entries, we use the double Hadamard parametrization $x = z_u \odot z_u - z_v \odot z_v$ and transform (C$_4$) to a nonconvex problem:

$$(z_u^\star, z_v^\star) \in \operatorname*{argmin}_{(z_u, z_v) \in \mathcal{B}_{2n}^2} f(z_u \odot z_u - z_v \odot z_v) \tag{NC$_4$}$$

Similar landscape results hold here.

Theorem 9. *Suppose f is convex and $(z_u^\star, z_v^\star)$ is any stationary point of* (NC$_4$)*. Then $(z_u^\star, z_v^\star)$ is a second-order stationary point of* (NC$_4$) *if and only if $z_u^\star \odot z_u^\star - z_v^\star \odot z_v^\star$ is a global minimizer of* (C$_4$)*.*

7.2 Extension to nonconvex functions

We can extend our framework to nonconvex objective functions f in Problems (C$_1$)–(C$_4$) as we only use the convexity of f in showing:

(P$_1$) The Hessian $\nabla^2 f(x^\star)$ of Problems (C$_1$)–(C$_4$) evaluated at the KKT points $x^\star$ is PSD;[4]

(P$_2$) Any KKT point of Problems (C$_1$)–(C$_4$) is a global minimizer.

Thus we can relax the convexity assumption of f and obtain more general results by assuming (P$_1$) and possibly (P$_2$) hold directly. For simplicity, let $h(z)$ denote the Hadamard parametrization given by $h(z) = z \odot z$ for (C$_1$)–(C$_3$) and $h(z) = z_u \odot z_u - z_v \odot z_v,$ for (C$_4$).

Theorem 10. *Under Assumption* (P$_1$)*, any stationary point $z^\star$ of Problems* (NC$_1$)–(NC$_4$) *is a second-order stationary point if and only if $h(z^\star)$ is a KKT point of Problems* (C$_1$)–(C$_4$)*.*

Note that this general result applies to a broader family of objective functions, which are not necessarily convex, such as quasi-convex functions and ℓ_p norm with $p \in (0, 1)$. If we further assume (P$_2$) then Theorem 10 will be stronger: any stationary point $z^\star$ of Problems (NC$_1$)–(NC$_4$) is either a strict saddle point of Problems (NC$_1$)–(NC$_4$) or $h(z^\star)$ is a global minimizer of Problems (C$_1$)–(C$_4$).

[4] See eqs. (16), (38), (47) and (56).

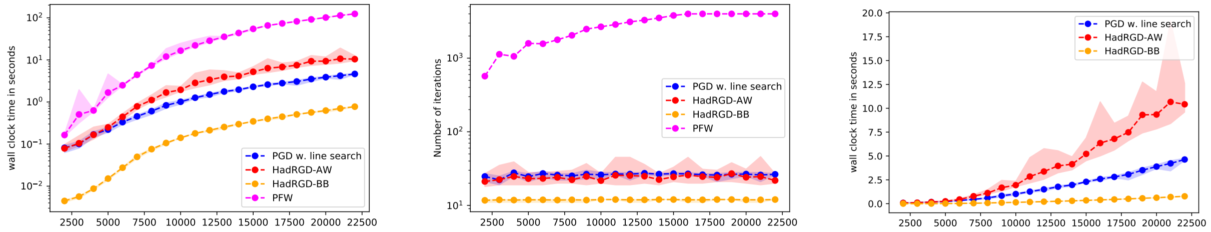

Figure 3: Solving underdetermined least squares with $x_{\text{true}} \in \text{int}(\Delta_n)$ and a target solution accuracy of 10^{-8}. **Left:** Time required. **Center:** Number of iterations required. **Right:** Time required, for the three fastest algorithms, not in log scale. All results are averaged over ten trials and the shading denotes the min-max range.

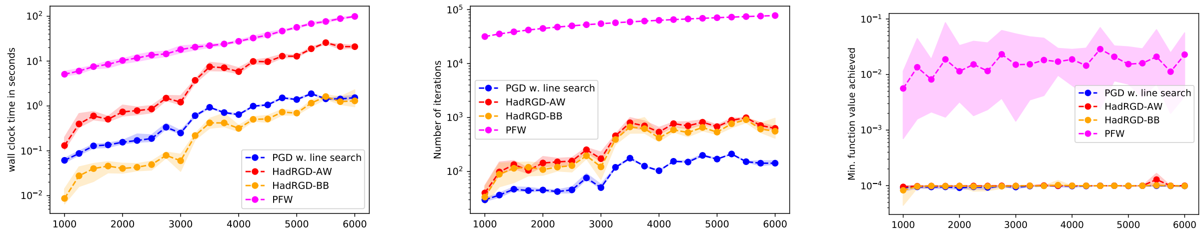

Figure 4: Solving underdetermined least squares with x_{true} on the boundary of Δ_n and a target solution accuracy of 10^{-4}. **Left:** Time required. **Center:** Number of iterations required. **Right:** Final solution accuracy; PFW struggles to reach the target accuracy. All results are averaged over ten trials and the shading denotes the min-max range.

8 Experiments

We consider simplex-constrained least squares[5]:

$$x^\star = \operatorname*{argmin}_{x \in \Delta_n} \left\{ f(x) = \|Ax - b\|_2^2 \right\} \tag{23}$$

where $A \in \mathbb{R}^{m \times n}$. We focus on this test problem as it is a key component of algorithms for applications as diverse as portfolio optimization (Bomze, 2002), archetypal analysis (Bauckhage et al., 2015) and hyperspectral unmixing (Condat, 2016). We take $m = 0.1n$ so that f is convex but *not strongly convex*. We take $b = Ax_{\text{true}}$ where $x_{\text{true}} \in \Delta_n$ is randomly selected in two qualitatively different ways: (i) x_{true} is sampled uniformly at random from the interior of Δ_n and (ii) x_{true} is the projection of a Gaussian random vector to Δ_n (and hence lies on the boundary of Δ_n).

We considered the following benchmark algorithms: Entropic Mirror Descent Algorithm (EMDA) (Ben-Tal et al., 2001; Beck and Teboulle, 2003); Pairwise Frank-Wolfe (PFW) (Pedregosa et al., 2020), both with and without linesearch; and Projected Gradient Descent (PGD) on Δ_n, with and without linesearch. See Appendix G for more details and supplementary for codes.

We found EMDA and PFW (without line search) to be non-competitive in both cases (see Appendix H), hence here we only present results for PGD (with line search), HadGrad-AW, HadGrad-BB and PFW (with line search). The results for case (i) are presented in Figure 3 while the results for case (ii) are presented in Figure 4. We record, as a function of n, the wall-clock time required to find an ε-solution ($\varepsilon = 10^{-8}$ for

[5] All code is available at https://github.com/DanielMckenzie/HadRGD

case (i) and $\varepsilon = 10^{-4}$ for case (ii)) or until the maximum number of iterations is reached (1000 for PGD, HadGrad-AW, HadGrad-BB and $1000\sqrt{n}$ for PFW).

Our experiments confirm HadGrad-AW enjoys a linear convergence rate when $x_{\text{true}} \in \text{int}(\Delta_n)$, as predicted by Theorem 6, see Figure 8. A significant amount of the run-time of HadGrad-AW is spent on computing points along the geodesic $\exp_z(\alpha v)$. This can be ameliorated by using a *retraction* within the line search, instead of a geodesic Boumal (2020). We leave this for future work. From Figure 3 it is clear that HadGrad-BB is the fastest algorithm for problem (23), converging an order of magnitude than the others, both in terms of number of iterations and wall-clock time.

9 Conclusion

In this paper we presented a new framework for transforming a non-smooth constraint set to the unit sphere or ball. We showed that this transformed problem may sometimes be solved much faster using techniques from Riemannian optimization. The superiority of our algorithms is particularly pronounced when: (i) the objective function f is convex but not strongly convex, and (ii) the cost of evaluating $\nabla f(x)$ is equal to or higher than the cost of projection on to Δ_n. When these two conditions are met we recommend practitoners consider using either HadRGD-AW (if a guaranteed convergence rate is required) or HadRGD-BB. Finally, we note that our approach is generic in the sense that it does not specify a particular optimization algorithm to use. Thus, future work could combine our Hadamard reparametrization trick with proximal or sub-gradient methods for non-smooth f, or SGD-style methods for finite sum objectives: $f(x) = \sum_{i=1}^{n} f_i(x)$.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Ali, A., Kolter, J. Z., and Tibshirani, R. J. (2019). A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1370–1378. PMLR.

Bauckhage, C., Kersting, K., Hoppe, F., and Thurau, C. (2015). Archetypal analysis as an autoencoder. In *Workshop New Challenges in Neural Computation*, page 8.

Bauer, F., Pereverzev, S., and Rosasco, L. (2007). On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

Ben-Tal, A., Margalit, T., and Nemirovski, A. (2001). The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1):79–108.

Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

Blum, M., Floyd, R. W., Pratt, V. R., Rivest, R. L., Tarjan, R. E., et al. (1973). Time bounds for selection. *J. Comput. Syst. Sci.*, 7(4):448–461.

Bomze, I. M. (2002). Regularity versus degeneracy in dynamics, games, and optimization: A unified approach to different aspects. *SIAM review*, 44(3):394–414.

Boumal, N. (2020). An introduction to optimization on smooth manifolds. *Available online, Aug*.

Bühlmann, P. and Yu, B. (2003). Boosting with the ℓ_2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339.

Bunea, F., Tsybakov, A. B., Wegkamp, M. H., and Barbu, A. (2010). Spades and mixture models. *The Annals of Statistics*, 38(4):2525–2558.

Chen, Y. and Ye, X. (2011). Projection onto a simplex. *arXiv preprint arXiv:1101.6081*.

Clarkson, K. L. (2010). Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):1–30.

Condat, L. (2016). Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming, Series A*, 158(1):575–585.

Criscitiello, C. and Boumal, N. (2019). Efficiently escaping saddle points on manifolds. *Advances in Neural Information Processing Systems*, 32:5987–5997.

De Klerk, E. (2008). The complexity of optimizing over a simplex, hypercube or sphere: A short survey. *Central European Journal of Operations Research*, 16(2):111–125.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279.

Fabian Pedregosa, Geoffrey Negiar, G. D. (2020). COPT: Composite OPTimization in python.

Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR.

Jaggi, M. (2013). Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR.

Keshava, N. (2003). A survey of spectral unmixing algorithms. *Lincoln laboratory journal*, 14(1):55–78.

Koh, K., Kim, S.-J., and Boyd, S. (2007). An interior-point method for large-scale ℓ_1-regularized logistic regression. *Journal of Machine learning research*, 8(Jul):1519–1555.

Lacoste-Julien, S. and Jaggi, M. (2015). On the global linear convergence of Frank-Wolfe optimization variants. In *NIPS 2015-Advances in Neural Information Processing Systems 28*.

Limmer, S. and Stańczak, S. (2018). A neural architecture for Bayesian compressive sensing over the simplex via Laplace techniques. *IEEE Transactions on Signal Processing*, 66(22):6002–6015.

Luenberger, D. G. (1972). The gradient projection method along geodesics. *Management Science*, 18(11):620–631.

Nemirovski, A. (2000). Ecole d'ete de probabilites de saint-flour XXVIII, chapter'topics in non-parametric statistics'.

Nesterov, Y. and Polyak, B. T. (2006). Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205.

Pedregosa, F., Negiar, G., Askari, A., and Jaggi, M. (2020). Linearly convergent Frank-Wolfe with backtracking line-search. In *International Conference on Artificial Intelligence and Statistics*, pages 1–10. PMLR.

Perez, G., Barlaud, M., Fillatre, L., and Régin, J.-C. (2020). A filtered bucket-clustering method for projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 182(1):445–464.

Shalev-Shwartz, S. and Singer, Y. (2006). Efficient learning of label ranking by soft projections onto polyhedra.

Sun, J., Qu, Q., and Wright, J. (2015). Complete dictionary recovery over the sphere. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 407–410. IEEE.

Vaskevicius, T., Kanade, V., and Rebeschini, P. (2019). Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32:2972–2983.

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434.

Yang, Y. (2007). Globally convergent optimization algorithms on Riemannian manifolds: Uniform framework for unconstrained and constrained optimization. *Journal of Optimization Theory and Applications*, 132(2):245–265.

Zeeman, E. C. (1980). Population dynamics from game theory. In *Global theory of dynamical systems*, pages 471–497. Springer.

Zhang, H. and Hager, W. W. (2004). A nonmonotone line search technique and its application to unconstrained optimization. *SIAM journal on Optimization*, 14(4):1043–1056.

Zhang, J. and Zhang, S. (2018). A cubic regularized Newton's method over riemannian manifolds. *arXiv preprint arXiv:1805.05565*.

Zhao, P., Yang, Y., and He, Q.-C. (2019). Implicit regularization via Hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*.

A Proof of Hadamard Calculus

(H1) $1_n \odot z = z$ where 1_n is the all-one vector in $\mathbb{R}^n$

Proof. This follows from the definition of Hadamard product. ■

(H2) If $d \odot z \odot z = 0$ then $d \odot z = 0$

Proof. We can multiply d on both sides of left-hand-side equation: $d \odot d \odot z \odot z = 0$, which implies $(d \odot z) \odot (d \odot z) = 0$. So, Therefore, $d \odot z = 0$. ■

(H3) $\mathrm{diag}(z)d = z \odot d$ and $d^\top \mathrm{diag}(z)d = \langle d, z \odot d\rangle = \langle z, d \odot d\rangle$

Proof. The first line is because $[\mathrm{diag}(z)d]_k = z_k d_k = [z \odot d]_k$. The second line follows by multiplying $d^\top$ on both sides and $\langle d, z \odot d\rangle = \sum_k d_k z_k d_k = \sum_k z_k d_k^2 = \langle z, d \odot d\rangle$. ■

(H4) $\|d \odot z\|_2 \le \|d\|_2 \|z\|_\infty$

Proof. Note that

$$\|d \odot z\|_2 = \sqrt{\sum_k d_k^2 z_k^2} \le \sqrt{\sum_k d_k^2 z_{\max}^2} = z_{\max}\sqrt{\sum_k d_k^2} = \|d\|_2 \|z\|_\infty$$

where $z_{\max}^2 := \max_k z_k^2$. ■

(H5) $\nabla g(z) = 2\nabla f(x) \odot z$ and $\nabla^2 g(z) = 2\mathrm{diag}(\nabla f(x)) + 4\mathrm{diag}(z)\nabla^2 f(x)\mathrm{diag}(z)$.

Proof. One one hand, the Taylor expansion of g around z for some arbitrary small direction d is given by

$$g(z+d) = g(z) + d^\top \nabla g(z) + \frac{1}{2} d^\top \nabla^2 g(z) d + o(\|d\|_2^2) \tag{24}$$

On the other hand, we have

$$\begin{aligned} g(z+d) &= f((z+d) \odot (z+d)) = f(z \odot z + 2z \odot d + d \odot d) \\ &= f(z \odot z) + (2z \odot d + d \odot d)^\top \nabla f(z \odot z) + \frac{1}{2}(2z \odot d)^\top \nabla^2 f(z \odot z)(2z \odot d) + o(\|d\|^2) \\ &= g(z) + 2d^\top(\nabla f(x) \odot z) + (d \odot d)^\top \nabla f(x) + \frac{4}{2} d^\top [\mathrm{diag}(z)\nabla^2 f(x)\,\mathrm{diag}(z)]d + o(\|d\|^2) \\ &= g(z) + 2d^\top(\nabla f(x) \odot z) + d^\top \,\mathrm{diag}(\nabla f(x))d + 2d^\top[\mathrm{diag}(z)\nabla^2 f(x)\,\mathrm{diag}(z)]d + o(\|d\|^2) \end{aligned} \tag{25}$$

Combining and rearranging (24) and (25), we complete the proof.

■

B Proof of Lemma 3.1

Lemma 3.1. *Suppose $f(x)$ is L-Lipschitz differentiable. Then g is $\tilde{L}$-Lipschitz differentiable with $\tilde{L} = 4L + 2M$ where $M = \max_{x \in \Delta_n} \|\nabla f(x)\|_2$.*

Proof. Recalling $\nabla_z g(z) = 2\nabla_x f(z \odot z) \odot z$ we compute

$$\begin{aligned}
\|\nabla g(z_1) - \nabla g(z_2)\|_2 &= 2\|\nabla_x f(z_1 \odot z_1) \odot z_1 - \nabla_x f(z_2 \odot z_2) \odot z_2\|_2 \\
&= 2\|\nabla_x f(z_1 \odot z_1) \odot z_1 - \nabla_x f(z_2 \odot z_2) \odot z_1 + \nabla_x f(z_2 \odot z_2) \odot z_1 - \nabla_x f(z_2 \odot z_2) \odot z_2\|_2 \\
&\leq 2\| \left(\nabla_x f(z_1 \odot z_1) - \nabla_x f(z_2 \odot z_2)\right) \odot z_1\|_2 + 2\|\nabla_x f(z_2 \odot z_2) \odot \left(z_1 - z_2\right)\|_2 \\
&\leq 2\|\nabla_x f(z_1 \odot z_1) - \nabla_x f(z_2 \odot z_2)\|_2 \|z_1\|_\infty + 2\|\nabla_x f(z_2 \odot z_2)\|_\infty \|z_1 - z_2\|_2 \\
&\leq 2\left(L\|z_1 \odot z_1 - z_2 \odot z_2\|_2\right)(1) + 2M\|z_1 - z_2\|_2 \\
&\leq 2L\left(2\|z_1 - z_2\|_2\right) + 2M\|z_1 - z_2\|_2 \\
&= \left(4L + 2M\right)\|z_1 - z_2\|_2
\end{aligned}$$

where the second inequality follows from (H4) while the last inequality follows from Lemma B.1. ■

Lemma B.1. *If $z_1, z_2 \in \mathcal{S}_n$ then $\|z_1 \odot z_1 - z_2 \odot z_2\|_2 \leq 2\|z_1 - z_2\|_2$.*

Proof.

$$\begin{aligned}
\|z_1 \odot z_1 - z_2 \odot z_2\|_2 &= \|z_1 \odot z_1 - z_1 \odot z_2 + z_1 \odot z_2 - z_2 \odot z_2\|_2 \\
&\leq \|z_1 \odot (z_1 - z_2)\|_2 + \|(z_1 - z_2) \odot z_2\|_2 \\
&\leq \|z_1\|_\infty \|z_1 - z_2\|_2 + \|z_1 - z_2\|_2 \|z_2\|_\infty \\
&\leq (1)\|z_1 - z_2\|_2 + \|z_1 - z_2\|_2 (1) = 2\|z_1 - z_2\|_2
\end{aligned}$$

where the second inequality follows from (H4). ■

C Proof of Theorem 2

Theorem 2. *Suppose $x^\star$ is a global minimizer of (C_1) with $x^\star \in \operatorname{int}(\Delta_n)$ and $\nabla^2 f(x^\star)$ PD on $\mathcal{U} := \{x \in \mathbb{R}^n : \sum_i x_i = 0\}$. Then all the coresponding Hadamard parametrizations $z_\star \in \mathcal{Z}^\star$ are nondegenerate stationary points of (NC_1).*

Proof. Using (15) and (11), we can show $z^\star$ is a nondegenerate stationary point of Problem (NC_1) if

$$(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) > 0, \ \forall d \perp z^\star, d \neq 0 \tag{26}$$

Note that:

$$d \perp z^\star \Rightarrow \sum_i d_i z_i^\star = 0 \Rightarrow u := z^\star \odot d \in \mathcal{U} \tag{27}$$

It is easy to recognize that

$$(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) = u^\top \nabla^2 f(x^\star) u \quad \text{with } u \in \mathcal{U}$$

By assumption, $u^\top \nabla^2 f(x^\star) u > 0$ as long as $u \neq 0$. As $z^\star \in \operatorname{int}(\Delta_n)$ we know $z_i^\star > 0$ for all i. Thus $u := z^\star \odot d = 0$ if and only if $d = 0$, and so (26) holds. ■

D Proof of Theorem 4

We begin with a lemma.

Lemma D.1. *Suppose g is twice continuously differentiable and has at least one second-order stationary point. Then, for any $\delta > 0$ there exists an $\epsilon_\delta > 0$ such that:*

$$z \text{ is an } \epsilon_\delta\text{-second-order stationary point } \Rightarrow \|z - z^\star\| < \delta \tag{28}$$

where $z^\star$ is a second-order stationary point.

Proof. Fix $\delta > 0$ and suppose to the contrary that no such ϵ_δ exists. Then, for any $\epsilon^k := 1/k$ we may find an ϵ^k-second-order stationary point z^k such that:

$$\|z^k - z^\star\| \geq \delta \text{ for all second-order stationary points } z^\star. \tag{29}$$

As $\mathcal{S}_n$ is compact, passing to a subsequence if necessary we may assume z^k converges: $\lim_{k\to\infty} z^k = \tilde{z} \in \mathcal{S}_n$. By continuity, $\tilde{z}$ is a second-order stationary point:

$$\begin{aligned}
&\|\operatorname{grad} g(\tilde{z})\| = \|\operatorname{grad} g(\lim_{k\to\infty} z^k)\| = \lim_{k\to\infty} \|\operatorname{grad} g(z^k)\| \leq \lim_{k\to\infty} \epsilon^k = 0 \\
&\lambda_{\min}(\operatorname{Hess} g(\tilde{z})) = \lambda_{\min}(\operatorname{Hess} g(\lim_{k\to\infty} z^k)) = \lim_{k\to\infty} \lambda_{\min}(\operatorname{Hess} g(z^k)) \geq \lim_{k\to\infty} \epsilon^k = 0
\end{aligned}$$

As $\tilde{z} = \lim_{k\to\infty} z^k$ this contradicts (29). □

Theorem 4. *Suppose $x^\star \in \operatorname{int}(\Delta_n)$ and $\nabla^2 f(x^\star)$ positive definite (PD) for every global minimizer of (C_1). Then HadPRGD finds an ϵ-optimal solution to (C_1) w.h.p. in $\mathcal{O}((\log n)^4/\epsilon)$ iterations, for ϵ small enough.*

Proof. Fix z_0. From Theorem 3 and Lemma D.1 we infer that if $\{z_k\}$ is the sequence of iterates generated by HadPRGD then $z_k \to z^\star$ for some second-order stationary point $z^\star$. From Theorem 1 it follows $x^\star := z^\star \odot z^\star$ is a global minimizer of f, and so by assumption $x^\star \in \operatorname{int}(\Delta_n)$ and $\nabla^2 f(x^\star)$ is positive definite.

From Theorem 2 it follows $z^\star$ is a nondegenerate stationary point of g; equivalently $\operatorname{Hess} g(z^\star)$ is positive definite. By continuity there exists a geodesic ball $B(z^\star, \delta)$ upon which $\operatorname{Hess} g(z)$ is positive definite, *i.e.* g restricted to $B(z^\star, \delta)$ is τ strongly convex for some $\tau > 0$ and so satisfies the PL inequality: $g(z) - g(z^\star) \leq \frac{1}{2\tau}\|\operatorname{grad} g(z)\|^2$.

From Lemma D.1 there exists $\epsilon_\delta > 0$ such that if z is an ϵ_δ-second-order stationary point then $z \in B(z^\star, \delta)$. So, suppose $\epsilon > 0$ is small enough that $\sqrt{\epsilon} < \epsilon_\delta$. By Theorem 3 HadPRGD finds a $\sqrt{\epsilon}$-second-order stationary point, call it z_K, in $\mathcal{O}\left((\log n)^4/(\sqrt{\epsilon})^2\right) = \mathcal{O}\left((\log n)^4/\epsilon\right)$ iterations (w.h.p). As $\sqrt{\epsilon} < \epsilon_\delta$ we know z_K is also an ϵ_δ-second-order stationary point hence $z_K \in B(z^\star, \delta)$. Appealing to the PL inequality and letting $x_K = z_K \odot z_K$:

$$f(x_K) - f(x^\star) = g(z_K) - g(z^\star) \leq \frac{1}{2\tau}\|\operatorname{grad} g(z_K)\|^2 \leq \frac{1}{2\tau}\left(\sqrt{\epsilon}\right)^2 \leq \epsilon \tag{30}$$

□

E Algorithms

Recall $\alpha^\star$ satisfies the (Riemannian) Armijo-Wolfe conditions along the geodesic traced out by $\exp_z(\alpha v)$ if

$$g\left(\exp_z(\alpha^\star v)\right) \leq g\left(z\right) + \rho_1 \alpha^\star \langle \operatorname{grad} g(z), v\rangle \tag{31a}$$

$$g'\left(\exp_z(\alpha^\star v)\right) \geq \rho_2 \langle \operatorname{grad} g(z), v\rangle \tag{31b}$$

When $v = -\operatorname{grad} g(z)$ this simplifies to:

$$g\left(\exp_z(\alpha^\star v)\right) \leq g\left(z\right) - \rho_1 \alpha^\star \|\operatorname{grad} g(z)\|_2^2 \tag{32a}$$

$$g'\left(\exp_z(\alpha^\star v)\right) \geq \rho_2 \|\operatorname{grad} g(z)\|_2^2 \tag{32b}$$

We use this version of the Armijo-Wolfe conditions in Algorithm 4, with the notation $\nabla_k = \operatorname{grad} g(z_k)$.

Algorithm 2 HadPRGD for (C_1)

```
Input: $x_0 \in \Delta_n$: initial point, $\alpha$: step size, $K$: number of iterations, $f$: original objective function
$z_0 = \sqrt{x_0}$ {(Defined componentwise)}
$g(z) := f(z \odot z)$
for k=1,..., K do
  if $\| \operatorname{grad} g(z_k) \|_2 > \varepsilon$ then
    $z_{k+1} = \exp_{x_k}(-\alpha \operatorname{grad} f(x_k))$
  else
    $\xi \sim \text{Uniform}(B_{z_k,r}(0))$
    $s_0 = \eta\xi$
    $z_{k+\mathcal{T}} \leftarrow \text{TangentSpaceSteps}(z_k, s_0, \eta, b, \mathcal{T})$ {(See Algorithm 3)}
  end if
end for
Return $x_K = z_K \odot z_K$
```

Algorithm 3 TangentSpaceSteps

```
Input: $z, s_0, \alpha, \eta, b, \mathcal{T}$.
for $j = 1, \ldots, \mathcal{T}$ do
  $s_{j+1} = s_j - \eta \operatorname{grad} f_z(s_j)$
  if $\|s_{j+1}\|_2 \geq b$ then
    $s_{\mathcal{T}} = s_j - \alpha\eta\nabla \hat{f}_z(s_j)$
    Break.
  end if
end for
Return $\text{Proj}_z(s_{\mathcal{T}})$
Return $x_K = z_K \odot z_K$
```

Algorithm 4 HadRGD-AW for (C_1)

Input: $x_0 \in \Delta_n$: initial point, α_{def}: default step size, β: decay factor, ρ_1: Armijo condition tolerance, ρ_2: Wolfe condition tolerance, K: number of iterations, f: original objective function

```
$z_0 = \sqrt{x_0}$ {(Defined componentwise)}
$g(z) := f(z \odot z)$
for $k = 1, \ldots, K$ do
  $\nabla_k \leftarrow \operatorname{grad} g(z_k)$
  $m = 0$
  ArmijoFlag=False
  WolfeFlag=False
  while ArmijoFlag=False and WolfeFlag=False and $m \le 25$ do
    $\alpha \leftarrow \alpha_{\text{def}}\beta^m$
    $m \leftarrow m + 1$
    if $g\left(\exp_{z_k}(\alpha\nabla_k)\right) \le g(z) - \rho_1\alpha\|\nabla_k\|_2^2$ then
      ArmijoFlag=True
    end if
    if $g'\left(\exp_{z_k}(\alpha^\star v)\right) \ge -\rho_2\|\nabla_k\|_2^2$ then
      WolfeFlag=True
    end if
  end while
  $z_{k+1} \leftarrow \exp_{z_k}(-\alpha\nabla_k)$
end for
Return $x_K = z_K \odot z_K$
```

Algorithm 5 HadRGD-BB for (C_1)

Input: $x_0 \in \Delta_n$: initial point, α_{def}: default step size, δ: decay factor, η: moving average factor, ρ_1: tolerance factor, K: number of iterations, f: original objective function

```
$z_0 = \sqrt{x_0}$ {(Defined componentwise)}
$g(z) := f(z \odot z)$
$\alpha \leftarrow \alpha_{\text{def}}$
for $k = 1, \ldots, K$ do
  $\nabla_k \leftarrow \operatorname{grad} g(z_k)$
  while $g(\exp_{z_k}(-\alpha\nabla_k) \ge C_k - \rho_1\alpha\|\nabla_k\|_2^2$ do
    $\alpha \leftarrow \alpha\delta$
  end while
  $z_{k+1} \leftarrow \exp_{z_k}(-\alpha\nabla_k)$
  $Q_{k+1} \leftarrow \eta Q_k + 1$
  $C_{k+1} \leftarrow \left(\eta Q_k C_k + g(z_{k+1})\right)/Q_{k+1}$
  $s_{k+1} \leftarrow z_{k+1} - z_k$
  $y_{k+1} \leftarrow \operatorname{grad} g(z_{k+1}) - \operatorname{grad} g(z_k)$
  $\alpha_{k+1}^{\text{BB}} \leftarrow \frac{\|s_{k+1}\|_2^2}{|\langle s_{k+1}, y_{k+1}\rangle|}$
  $\alpha \leftarrow \max\left\{\min\left\{\alpha_{k+1}^{\text{BB}}, 30\right\} 10^{-10}\right\}$
end for
Return $x_K = z_K \odot z_K$
```

Algorithm 6 PGD with line search for (C$_1$)

Input: $x_0 \in \Delta_n$: initial point, $P_{\Delta_n}(\cdot)$: A callable projection algorithm, s: step size, β: decay factor, ρ_1: Armijo condition tolerance K: number of iterations, f: original objective function

```
for $k = 1, \ldots, K$ do
    $\nabla_k \leftarrow \nabla f(x_k)$
    $\bar{x} \leftarrow P_{\Delta_n}(x_k - s\nabla_k)$
    $m = 0$
    ArmijoFlag= False
    while ArmijoFlag= False and $m \leq 25$ do
        $\alpha \leftarrow \beta^m$
        $x_{\text{new}} \leftarrow x_k + \alpha(\bar{x} - x_k)$
        if $f(x_k) - f(x_{\text{new}}) \geq -\rho_1 \langle \nabla_k, \bar{x} - x_k \rangle$ then
            ArmijoFlag= True
        end if
    end while
    $x_{k+1} \leftarrow x_{\text{new}}$
end for
```

Return $x_K = z_K \odot z_K$

F Proofs of Landscape Analysis

F.1 The Unit Simplex

Recall the original problem is

$$x^\star \in \operatorname*{argmin}_{x \in \blacktriangle_n} f(x) \tag{C$_2$}$$

and the Hadamard parameterized problem is

$$z^\star \in \operatorname*{argmin}_{z \in \mathcal{B}_n^2} f(z \odot z) \tag{NC$_2$}$$

Theorem 7. *Suppose f is convex and $z^\star$ is any stationary point $z^\star$ of (NC$_2$). Then $z^\star$ is a second-order stationary point of (NC$_2$) if and only if $z^\star \odot z^\star$ is a global minimizer of (C$_2$).*

KKT conditions of Problem (C$_2$) The Lagrangian of Problem (C$_2$) is

$$\mathcal{L}_C(x, \lambda, \beta) = f(x) - \lambda(1_n^\top x - 1) + \beta^\top x \tag{33}$$

Since Problem (C$_2$) is convex, the global optimality conditions of Problem (C$_2$) are given by the KKT conditions that there exist $\lambda^\star \in \mathbb{R}$ and $\beta^\star \in \mathbb{R}^n$ such that

$$\nabla f(x^\star) = \lambda^\star 1 + \beta^\star \tag{34a}$$

$$x^\star \geq 0 \tag{34b}$$

$$\beta^\star \geq 0 \tag{34c}$$

$$x^\star \odot \beta^\star = 0 \tag{34d}$$

$$1_n^\top x^\star \leq 1 \tag{34e}$$

$$\lambda^\star \leq 0 \tag{34f}$$

$$(1_n^\top x^\star - 1)\lambda^\star = 0 \tag{34g}$$

First-order optimality conditions of Problem (NC$_2$) The Lagrangian of Problem (NC$_2$) is

$$\mathcal{L}_N(z, \lambda_N) = f(z \odot z) - \lambda_N(\|z\|_2^2 - 1) \tag{35}$$

Then the first-order optimality conditions of Problem (NC$_2$) are given by that there exist $\lambda_N^\star \in \mathbb{R}$ such that

$$\begin{aligned} \nabla f(z^\star \odot z^\star) \odot z^\star &= \lambda_N^\star z^\star && \text{(36a)} \\ \|z^\star\|^2 &\le 1 && \text{(36b)} \\ \lambda_N^\star &\le 0 && \text{(36c)} \\ (\|z^\star\|^2 - 1))\lambda_N^\star &= 0 && \text{(36d)} \end{aligned}$$

Second-order optimality conditions of Problem (NC$_2$) The second-order optimality conditions of Problem (NC$_2$) are given by that there exist $\lambda_N^\star \in \mathbb{R}$ such that $(z^\star, \lambda_N^\star)$ satisfy (36) and

$$\begin{aligned} & d^\top \left[\nabla^2 \mathcal{L}_N(z^\star, \lambda_N^\star)\right] d \ge 0, \ \forall d \\ \iff & d^\top \left[2\mathrm{diag}(\nabla f(z^\star \odot z^\star)) + 4\mathrm{diag}(z^\star)\nabla^2 f(z^\star \odot z^\star)\mathrm{diag}(z^\star) - 2\lambda_N^\star I_n\right] d \ge 0, \ \forall d \\ \iff & 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) + 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda_N^\star \|d\|^2 \ge 0, \ \forall d \end{aligned} \tag{37}$$

Strict saddle point of Problem (NC$_2$) A stationary point $z^\star$ of (NC$_2$) not satisfying the second-order optimality conditions

Proof. The main proof consists of two steps. Denote $\mathcal{Z}^\star := \{z^\star : z^\star \odot z^\star = x^\star\}$ with $x^\star$ as any KKT point of (C$_2$).

1. **"If" part:** If $z^\star \in \mathcal{Z}^\star$, then $z^\star$ is a second order stationary point of Problem (NC$_2$). Suppose $z^\star \in \mathcal{Z}^\star$. Similar to before, firstly, we can show $z^\star$ satisfies the first-order optimality condition (36a) with $\lambda_N^\star = \lambda^\star$ by multiplying $z^\star$ on both sides of (34a). Secondly, by plugging $x^\star = z^\star \odot z^\star$ and $1_n^\top x^\star = \|z^\star\|^2$ to (34e)–(34g), we prove first-order optimality conditions (36b)–(36d) by choosing $\lambda_N^\star = \lambda^\star$. Hence, it remains to show the second-order optimality condition with $\lambda_N^\star = \lambda^\star$. Since $x^\star$ satisfies the optimality conditions (34) of Problem (C$_2$) and $x^\star = z^\star \odot z^\star$, we get

$$\begin{aligned} & \nabla f(x^\star) = \lambda^\star 1_n + \beta^\star, \beta \ge 0 \\ \iff & \mathrm{diag}(\nabla f(z^\star \odot z^\star)) - \lambda^\star I_n = \mathrm{diag}(\beta^\star) \succeq 0 \\ \iff & \langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - \lambda^\star \|d\|^2 \ge 0, \ \forall d \end{aligned}$$

   Then plugging this into the second-order optimality conditions (37) of Problem (NC$_2$) with $\lambda_N^\star = \lambda^\star$, we obtain

$$\begin{aligned} d^\top[\nabla^2 \mathcal{L}_N(z^\star)]d &= 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) + 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda^\star \|d\|^2 \\ &\ge 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) \\ &\ge 0, \ \forall d \end{aligned} \tag{38}$$

   where (38) follows from the convexity assumption of f. Therefore, $z^\star \in \mathcal{Z}^\star$ is a second-order stationary point of Problem (NC$_2$).

2. **"Only if" part:** $z^\star$ is a second order stationary point of Problem (NC$_2$) only if $z^\star \in \mathcal{Z}^\star$. For convenience, we show its contrapositive: If $z^\star \notin \mathcal{Z}^\star$, then $z^\star$ is a strict saddle point of Problem (NC$_2$).

Since $z^\star$ satisfies (36), there exists $\lambda_N^\star \in \mathbb{R}$ such that

$$\begin{aligned} \nabla f(z^\star \odot z^\star) \odot z^\star &= \lambda_N^\star z^\star \\ \|z^\star\|^2 &\leq 1 \\ \lambda^\star &\leq 0 \\ (\|z^\star\|^2 - 1))\lambda_N^\star &= 0 \end{aligned}$$

This implies that

$$\begin{aligned} [\nabla f(z^\star \odot z^\star)]_k &= \lambda_N^\star, \ \forall z_k^\star \neq 0 \\ z^\star \odot z^\star &\geq 0 \\ \|z^\star\|^2 &\leq 1 \\ \lambda_N^\star &\leq 0 \\ (1_n^\top (z^\star \odot z^\star) - 1))\lambda_N^\star &= 0 \end{aligned}$$

On the other hands, we recognize that the following equations in (34)

$$\begin{aligned} \nabla f(x^\star) &= \lambda^\star 1_n + \beta^\star \\ \beta^\star &\geq 0 \\ x^\star \odot \beta^\star &= 0 \end{aligned}$$

are equivalent to

$$\begin{cases} [\nabla f(x^\star)] = \lambda^\star, \ \forall x_k^\star \neq 0 \\ [\nabla f(x^\star)] \geq \lambda^\star, \ \forall x_k^\star = 0 \end{cases}$$

For the sake of contradiction, suppose further that

$$[\nabla f(z^\star \odot z^\star)] \geq \lambda_N^\star, \ \forall z_k^\star = 0$$

then $z^\star \odot z^\star$ satisfies (34) by choosing $\lambda^\star = \lambda_N^\star$, *i.e.* $z^\star \in \mathcal{Z}^\star$, which contradicts the assumption that $z^\star \notin \mathcal{Z}^\star$. Therefore, we must have

$$[\nabla f(z^\star \odot z^\star)]_{k^\star} < \lambda_N^\star \quad \text{for some } z_{k^\star}^\star = 0 \tag{39}$$

By constructing $d = e_{k^\star}$, where $\{e_i\}$ denotes the standard basis in $\mathbb{R}^n$, we have $d \odot z^\star = 0$. Then by direct computations we obtain

$$\begin{aligned} d^\top [\nabla^2 \mathcal{L}_N(z^\star, \lambda_N^\star)] d &= 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) + 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda_N^\star \|d\|^2 \\ &= 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda_N^\star \|d\|^2 \\ &= 2[\nabla f(z^\star \odot z^\star)]_{k^\star} - 2\lambda_N^\star \\ &\overset{(39)}{<} 0 \end{aligned}$$

Hence, $z^\star$ is a strict saddle point of Problem (NC$_2$).

■

F.2 The weighted probability simplex

Recall the original problem is

$$x^\star \in \operatorname*{argmin}_{x\in\Delta_n^a} f(x) \tag{C$_3$}$$

and the Hadamard parameterized problem is

$$z^\star \in \operatorname*{argmin}_{z\in\mathcal{B}_n^a} f(z\odot z) \tag{NC$_3$}$$

Theorem 8. *Suppose f is convex and $z^\star$ is any any stationary point $z^\star$ of (NC$_3$). Then $z^\star$ is a second-order stationary point of (NC$_3$) if and only if $z^\star\odot z^\star$ is a global minimizer of (C$_3$).*

KKT conditions of Problem (C$_3$) The Lagrangian of Problem (C$_3$) is

$$\mathcal{L}_C(x,\lambda,\beta) = f(x) - \lambda(a^\top x - 1) + \beta^\top x \tag{40}$$

Since Problem (C$_3$) is convex, the global optimality conditions are given by the KKT conditions that there exist $\lambda^\star\in\mathbb{R}$ and $\beta^\star\in\mathbb{R}^n$ such that

$$\begin{aligned} \nabla f(x^\star) &= \lambda^\star a + \beta^\star && \text{(41a)}\\ x^\star &\geq 0 && \text{(41b)}\\ \beta^\star &\geq 0 && \text{(41c)}\\ x^\star\odot\beta^\star &= 0 && \text{(41d)}\\ a^\top x^\star &= 1 && \text{(41e)} \end{aligned}$$

First-order optimality conditions of Problem (NC$_3$) The Lagrangian of Problem (NC$_3$) is

$$\mathcal{L}_N(z,\lambda_N) = f(z\odot z) - \lambda_N(\|z\|^2_{\mathrm{diag}(a)} - 1) \tag{42}$$

Then the first-order optimality conditions of Problem (NC$_3$) are given by that there exist $\lambda^\star\in\mathbb{R}$ such that

$$\begin{aligned} \nabla f(z^\star\odot z^\star)\odot z^\star &= \lambda_N^\star a\odot z^\star && \text{(43a)}\\ \|z^\star\|^2_{\mathrm{diag}(a)} &= 1 && \text{(43b)} \end{aligned}$$

Second-order optimality conditions of Problem (NC$_3$) The second-order optimality conditions of Problem (NC$_3$) are given by that there exist $\lambda_N^\star\in\mathbb{R}$ such that $(z^\star,\lambda_N^\star)$ satisfy (43a) and (43b) and

$$\begin{aligned} & d^\top\left[\nabla^2\mathcal{L}_N(z^\star,\lambda_N^\star)\right]d \geq 0\\ \Longleftrightarrow\ & d^\top\left[2\mathrm{diag}(\nabla f(z^\star\odot z^\star)) + 4\mathrm{diag}(z^\star)\nabla^2 f(z^\star\odot z^\star)\mathrm{diag}(z^\star) - 2\lambda_N^\star\mathrm{diag}(a)\right]d\geq 0\\ \Longleftrightarrow\ & 2\langle\nabla f(z^\star\odot z^\star), d\odot d\rangle + 4(z^\star\odot d)^\top\nabla^2 f(z^\star\odot z^\star)(z^\star\odot d) - 2\lambda_N^\star\|d\|^2_{\mathrm{diag}(a)}\geq 0,\ \forall d\perp z^\star \end{aligned} \tag{44}$$

Proof. The proof consists of three steps. Before proceeding, define $\mathcal{Z}^\star := \{z^\star : z^\star\odot z^\star = x^\star\}$ with $x^\star$ as any KKT point of (C$_3$).

1. **"If" part:** If $z^\star \in \mathcal{Z}^\star$, then $z^\star$ is a second order stationary point of Problem (NC$_3$). Suppose $z^\star \in \mathcal{Z}^\star$. We first show $z^\star \in \mathcal{Z}^\star$ satisfies the first-order optimality conditions (43a) and (43b). Multiplying both sides of the optimality condition (41a) by $z^\star$:

$$\nabla f(x^\star) \odot z^\star = \lambda^\star a \odot z^\star + \beta^\star \odot z^\star \tag{45}$$
$$\Longrightarrow f(z^\star \odot z^\star) \odot z^\star = \lambda^\star a \odot z^\star + \beta^\star \odot z^\star \tag{46}$$

   By complementary slackness (41d): $z^\star \odot z^\star \odot \beta^\star = 0$, and so $z^\star \odot \beta^\star = 0$ by (H2), thus (46) reduces to (43a) by choosing $\lambda_N^\star = \lambda^\star$. Also note (41e) is equivalent to (43b):

$$1 = a^\top x^\star = a^\top z^\star \odot z^\star = \|z^\star\|^2_{\mathrm{diag}(a)}$$

   It remains to show $z^\star \in \mathcal{Z}^\star$ satisfies the second-order optimality conditions (44) of Problem (NC$_3$) with $\lambda_N^\star = \lambda^\star$. Since $x^\star$ satisfies (41a) and (41c):

$$\begin{aligned}
&\nabla f(x^\star) = \lambda^\star a + \beta^\star, \beta^\star \geq 0\\
\iff &\mathrm{diag}(\nabla f(z^\star \odot z^\star)) - \lambda^\star \mathrm{diag}(a) = \mathrm{diag}(\beta^\star) \succeq 0\\
\iff &d^\top \left[\mathrm{diag}(\nabla f(z^\star \odot z^\star)) - \lambda^\star \mathrm{diag}(a)\right] d \geq 0\\
\overset{\text{(H3)}}{\iff} &\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - \lambda^\star \|d\|^2_{\mathrm{diag}(a)} \geq 0, \ \forall d
\end{aligned}$$

   Plugging this into (44) by choosing $\lambda_N^\star = \lambda^\star$, we get

$$\begin{aligned}
d^\top[\nabla^2 \mathcal{L}_N(z^\star)]d &= 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) + 2\langle \nabla f(z^\star \odot z^\star), d \odot d\rangle - 2\lambda^\star \|d\|^2_{\mathrm{diag}(a)}\\
&\geq 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d)\\
&\geq 0, \ \forall d
\end{aligned} \tag{47}$$

   where (47) follows from the convexity assumption of f.

2. **"Only if" part:** $z^\star$ is second order stationary points of Problem (NC$_3$) only if $z^\star \in \mathcal{Z}^\star$. For convenience, we show its contrapositive: If $z^\star \notin \mathcal{Z}^\star$, then $z^\star$ is a strict saddle point of Problem (NC$_3$). First, note that $z^\star$ satisfies (43b), we have

$$z^\star \odot z^\star \geq 0 \text{ and } a^\top (z^\star \odot z^\star) = 1$$

   That is, $z \odot z$ satisfies the optimality conditions (41b) and (41e) of Problem (C$_3$). Second, since $z^\star$ satisfies (43a), there exists $\lambda_N^\star \in \mathbb{R}$ such that

$$\begin{aligned}
&\nabla f(z^\star \odot z^\star) \odot z^\star = \lambda_N^\star a \odot z^\star\\
\Longrightarrow &[\nabla f(z^\star \odot z^\star)]_k = \lambda_N^\star a_k, \ \forall z_k^\star \neq 0
\end{aligned}$$

   On the other hand, we recognize that (41a),(41c) and (41d) are equivalent to:

$$\begin{cases}
[\nabla f(x^\star)]_k = \lambda^\star a_k, \ \forall x_k^\star \neq 0\\
[\nabla f(x^\star)]_k \geq \lambda^\star a_k, \ \forall x_k^\star = 0
\end{cases}$$

   For the sake of contradiction, let's further suppose

$$[\nabla f(z^\star \odot z^\star)]_k \geq \lambda_N^\star a_k, \ \forall z_k^\star = 0$$

   then $z^\star \odot z^\star$ satisfies the optimality conditions (41a),(41c) and (41d) by choosing

$$\begin{aligned}
&\lambda^\star = \lambda_N^\star\\
&\beta_k^\star = [\nabla f(z^\star \odot z^\star)]_k - \lambda^\star a_k \geq 0, \ \forall k
\end{aligned}$$

Consequently, $z^\star \odot z^\star$ satisfies all five optimality conditions of Problem (C$_3$), *i.e.* $z^\star \in \mathcal{Z}^\star$, which is a contradiction to the assumption $z^\star \notin \mathcal{Z}^\star$. So, we must have

$$[\nabla f(z^\star \odot z^\star)]_{k^\star} < \lambda^\star a_k \quad \text{for some } z^\star_{k^\star} = 0 \tag{48}$$

By constructing $d = e_{k^\star}$, where $\{e_i\}$ denotes the canonical basis in $\mathbb{R}^n$, we have $\langle d, z^\star \rangle = 0$ and $d \odot z^\star = 0$. By direct computation and (44) we obtain

$$\begin{aligned} d^\top [\nabla^2 \mathcal{L}_N(z^\star, \lambda^\star_N)] d &= 4(z^\star \odot d)^\top \nabla^2 f(z^\star \odot z^\star)(z^\star \odot d) + 2\langle \nabla f(z^\star \odot z^\star), d \odot d \rangle - 2\lambda^\star_N \|d\|^2_{\text{diag}(a)} \\ &= 0 + 2\langle \nabla f(z^\star \odot z^\star), d \odot d \rangle - 2\lambda^\star_N \|d\|^2_{\text{diag}(a)} \\ &= 2[\nabla f(z^\star \odot z^\star)]_{k^\star} - 2\lambda^\star_N a_{k^\star} \\ &\overset{(48)}{<} 0 \end{aligned}$$

Thus, $z^\star$ is a strict saddle point of Problem (NC$_3$).

■

F.3 The ℓ_1 norm ball

Recall the original problem is

$$x^\star \in \operatorname*{argmin}_{x \in \mathcal{B}^1_n} f(x) \tag{C$_4$}$$

and the Hadamard parameterized problem is

$$(z^\star_u, z^\star_v) \in \operatorname*{argmin}_{(z_u, z_v) \in \mathcal{B}^2_{2n}} f(z_u \odot z_u - z_v \odot z_v) \tag{NC$_4$}$$

Theorem 9. *Suppose f is convex and $(z^\star_u, z^\star_v)$ is any any stationary point of (NC$_4$). Then $(z^\star_u, z^\star_v)$ is a second-order stationary point of (NC$_4$) if and only if $z^\star_u \odot z^\star_u - z^\star_v \odot z^\star_v$ is a global minimizer of (C$_4$).*

KKT conditions of Problem (C$_4$) The Lagrangian of Problem (C$_4$) is

$$\mathcal{L}_C(x, \lambda) = f(x) - \lambda \left(\|x\|_1 - 1 \right). \tag{49}$$

Since Problem (C$_4$) is convex, the global optimality conditions are given by the KKT conditions that there exists $\lambda^\star \in \mathbb{R}$ such that:

$$\nabla f(x^\star) \in \lambda^\star \operatorname{sign}(x^\star) \tag{50a}$$

$$\lambda^\star \le 0 \tag{50b}$$

$$\|x^\star\|_1 \le 1 \tag{50c}$$

$$\lambda^\star \left(\|x^\star\|_1 - 1 \right) = 0 \tag{50d}$$

where

$$[\operatorname{sign}(x)]_k = \begin{cases} 1 & \text{if } x_k > 0 \\ -1 & \text{if } x_k < 0 \\ [-1, 1] & \text{if } x_k = 0 \end{cases}$$

First-order optimality conditions of Problem (NC$_4$) The Lagrangian of Problem (NC$_4$) is

$$\mathcal{L}_N(z_u, z_v, \lambda_N) = f(z_u \odot z_u - z_v \odot z_v) - \lambda_N \left(\|z_u\|_2^2 + \|z_v\|_2^2 - 1\right) \tag{51}$$

Thus we derive the first-order optimality conditions:

$$\begin{aligned} \nabla f(\widetilde{x}) \odot z_u^\star &= \lambda_N^\star z_u^\star && \text{(52a)} \\ -\nabla f(\widetilde{x}) \odot z_v^\star &= \lambda_N^\star z_v^\star && \text{(52b)} \\ \|z_u^\star\|_2^2 + \|z_v^\star\|_2^2 &\le 1 && \text{(52c)} \\ \lambda_N^\star &\le 0 && \text{(52d)} \\ \lambda_N^\star \left(\|z_u^\star\|_2^2 + \|z_v^\star\|_2^2 - 1\right) &= 0 && \text{(52e)} \end{aligned}$$

where we denote $\widetilde{x} := z_u^\star \odot z_u^\star - z_v^\star \odot z_v^\star$ to simplify notations.

Second-order optimality conditions of Problem (NC$_4$) First of all, let us compute the Hessian of $\mathcal{L}_N$:

$$\nabla^2 \mathcal{L}_N(z_u^\star, z_v^\star, \lambda_N^\star) = 2 \begin{bmatrix} \operatorname{diag}(\nabla f(\widetilde{x})) - \lambda_N^\star I_n & \\ & -\operatorname{diag}(\nabla f(\widetilde{x})) - \lambda_N^\star I_n \end{bmatrix} + 4 \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix} \nabla^2 f(\widetilde{x}) \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix}^\top \tag{53}$$

The second-order optimality conditions claim that the following matrix is positive semi-definite (PSD):

$$2 \begin{bmatrix} \operatorname{diag}(\nabla f(\widetilde{x})) - \lambda_N^\star I_n & \\ & -\operatorname{diag}(\nabla f(\widetilde{x})) - \lambda_N^\star I_n \end{bmatrix} + 4 \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix} \nabla^2 f(\widetilde{x}) \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix}^\top \succeq 0 \tag{54}$$

Proof. Denote

$$\mathcal{Z}^\star = \{(z_u^\star, z_v^\star) : z_u^\star \odot z_u^\star - z_v^\star \odot z_v^\star = x^\star, \lambda^\star z_u^\star \odot z_v^\star = 0\}$$

where $x^\star$ is any KKT point of (C$_4$) and $\lambda^\star$ is the Lagrangian dual variable in (50b). As usual, we break the proof two parts.

1. **"If" part:** If $(z_u^\star, z_v^\star) \in \mathcal{Z}^\star$, then $(z_u^\star, z_v^\star)$ is a second order stationary point of Problem (NC$_4$). Suppose $(z_u^\star, z_v^\star) \in \mathcal{Z}^\star$. We first show $(z_u^\star, z_v^\star) \in \mathcal{Z}^\star$ satisfies the first-order optimality conditions (52a)–(52e) for Problem (NC$_4$). First of all, since $(z_u^\star, z_v^\star) \in \mathcal{Z}^\star$, we have $\widetilde{x} = x^\star$. Then the optimality conditions (52c)–(52e) directly follow from (50b)–(50d) by choosing $\lambda_N^\star = \lambda^\star$. It suffices to show the optimality conditions (52a)–(52b) with $\lambda_N^\star = \lambda^\star$. Rewriting (50a) we obtain:

$$\begin{aligned} \nabla[f(x^\star)]_i &= \lambda^\star && \text{if } x_i^\star > 0 \\ \nabla[f(x^\star)]_i &= -\lambda^\star && \text{if } x_i^\star < 0 \\ \nabla[f(x^\star)]_i &\in [\lambda^\star - \lambda^\star] && \text{if } x_i^\star = 0 \end{aligned} \tag{55}$$

Case 1: $\lambda^\star = 0$. Then we have $\nabla f(x^\star) = 0$, *i.e.* $\nabla f(\widetilde{x}) = 0$. Take $\lambda_N^\star = \lambda^\star = 0$. Then it is trivial that (52a)–(52b) is true.

Case 2: $\lambda^\star \neq 0$. Then the above equation (55) implies the decoupling property

$$u^\star \odot v^\star = 0.$$

Therefore,

$$x_i^\star = \widetilde{x}_i = \begin{cases} [z_u^\star]_i{}^2 & \text{if } [z_u^\star]_i \neq 0 \\ -[z_v^\star]_i{}^2 & \text{if } [z_v^\star]_i \neq 0 \end{cases}$$

which further indicates that for all $(z_u^\star, z_v^\star) \in \mathcal{Z}^\star$:

$$\begin{aligned} \nabla[f(\widetilde{x})]_i &= \lambda^\star \quad \text{if } \widetilde{x}_i = x_i^\star > 0 \iff [z_u^\star]_i \neq 0, [z_v^\star]_i = 0 \\ \nabla[f(\widetilde{x})]_i &= -\lambda^\star \quad \text{if } \widetilde{x}_i = x_i^\star < 0 \iff [z_v^\star]_i \neq 0, [z_u^\star]_i = 0 \\ \nabla[f(\widetilde{x})]_i &\in [\lambda^\star, -\lambda^\star] \quad \text{if } \widetilde{x}_i = x_i^\star = 0 \iff [z_u^\star]_i = 0, [z_v^\star]_i = 0 \end{aligned}$$

Multiplying $\nabla f(\widetilde{x})$ by $z_u^\star$ and $z_v^\star$ respectively, and combining the above three, we then get the optimality conditions (52a) and (52b), respectively.

It remains to show $(z_u^\star, z_v^\star) \in \mathcal{Z}^\star$ satisfies the second-order optimality condition (54) with $\lambda_N^\star = \lambda^\star$. In this case, we still have $\widetilde{x} = x^\star$. Since $\nabla f(x^\star) \in \lambda^\star \operatorname{sign}(x^\star)$ by (50a), we must have $\pm \operatorname{diag}(\nabla f(x^\star)) - \lambda^\star I_n \succeq 0$ (recall that $\lambda^\star \leq 0$). Therefore, we have

$$\pm \operatorname{diag}(\nabla f(\widetilde{x})) - \lambda^\star I_n \succeq 0.$$

Using this, to show the second-order optimality condition (54):

$$2 \begin{bmatrix} \operatorname{diag}(\nabla f(\widetilde{x})) - \lambda_N^\star I_n & \\ & -\operatorname{diag}(\nabla f(\widetilde{x})) - \lambda_N^\star I_n \end{bmatrix} + 4 \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix} \nabla^2 f(\widetilde{x}) \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix}^\top \succeq 0$$

it suffices to show that

$$4 \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix} \nabla^2 f(\widetilde{x}) \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix}^\top \succeq 0 \tag{56}$$

which follows from the convexity assumption of f.

2. **"Only if" part:** $(z_u^\star, z_v^\star)$ is second order stationary points of Problem (NC$_4$) only if $(z_u^\star, z_v^\star) \in \mathcal{Z}^\star$. For convenience, we show its contrapositive: If $(z_u^\star, z_v^\star) \notin \mathcal{Z}^\star$, then $(z_u^\star, z_v^\star)$ is a strict saddle point of Problem (NC$_4$). First of all, the combination of (52a)–(52b) shows that

$$\nabla f(\widetilde{x}) \odot z_u^\star \odot z_v^\star = \lambda_N^\star z_u^\star \odot z_v^\star = -\lambda_N^\star z_u^\star \odot z_v^\star$$

implying that

$$\lambda_N^\star z_u^\star \odot z_v^\star = 0 \tag{57}$$

Therefore, together the assumption that $(z_u^\star, z_v^\star) \notin \mathcal{Z}^\star$, we must have

$$\widetilde{x} \neq x^\star \tag{58}$$

This is because, otherwise, we can choose $\lambda^\star = \lambda_N^\star$ to certify the global optimality of $\widetilde{x}$) in Problem (C$_4$), *i.e.* $(z_u^\star, z_v^\star) \in \mathcal{Z}^\star$, which is a contradiction.

To proceed, we further note that the optimality conditions (52c)–(52e) directly implies (50b)–(50d) by choosing $\lambda^\star = \lambda_N^\star$. This means that the suboptimality condition (58) implies that

$$\nabla f(\widetilde{x}) \notin \lambda_N^\star \operatorname{sign}(\widetilde{x}) \tag{59}$$

Also in view of (52a) and (52b), we have

$$\nabla f(\widetilde{x}) \odot z_u^\star = \lambda_N^\star z_u^\star \Longrightarrow [\nabla f(\widetilde{x})]_k = \lambda_N^\star \quad \text{if } [z_u^\star]_k \neq 0 \tag{60a}$$

$$-\nabla f(\widetilde{x}) \odot z_v^\star = \lambda_N^\star z_v^\star \Longrightarrow [\nabla f(\widetilde{x})]_k = -\lambda_N^\star \quad \text{if } [z_v^\star]_k \neq 0 \tag{60b}$$

In the following, we proceed case by case.

Case 1: $\lambda_N^\star = 0$. Then

$$[\nabla f(\widetilde{x})]_k = 0 \quad \text{if either } [z_u^\star]_k \neq 0 \text{ or } [z_v^\star]_k \neq 0$$

Note that we cannot have $\nabla f(\widetilde{x}) = 0$, since then $\lambda^\star = 0$ could certify the global optimality of $\widetilde{x}$ in Problem (C$_4$), which contradicts with the fact $(z_u^\star, z_v^\star) \notin \mathcal{Z}^\star$. Therefore, there must be some $k^\star$ such that

$$[\nabla f(\widetilde{x})]_{k^\star} \neq 0 \quad \text{for some } [z_u^\star]_{k^\star} = [z_v^\star]_{k^\star} = 0. \tag{61}$$

In this case, plugging in $\lambda_N^\star = 0$, we can further simplify the Hessian

$$\nabla^2 \mathcal{L}_N(z_u^\star, z_v^\star, 0) = 2 \begin{bmatrix} \operatorname{diag}(\nabla f(\widetilde{x})) & \\ & -\operatorname{diag}(\nabla f(\widetilde{x})) \end{bmatrix} + 4 \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix} \nabla^2 f(\widetilde{x}) \begin{bmatrix} \operatorname{diag}(z_u^\star) \\ -\operatorname{diag}(z_v^\star) \end{bmatrix}^\top$$

Then we can choose the direction vector as

$$(d_u, d_v) = \begin{cases} (e_k^\star, 0) & \text{if } [\nabla f(\widetilde{x})]_{k^\star} < 0 \\ (0, e_k^\star) & \text{if } [\nabla f(\widetilde{x})]_{k^\star} > 0 \end{cases}$$

so that we have

$$\begin{bmatrix} d_u^\top & d_v^\top \end{bmatrix} \nabla^2 \mathcal{L}_N(z_u^\star, z_v^\star, 0) \begin{bmatrix} d_u \\ d_v \end{bmatrix} = -2\,|[\nabla f(\widetilde{x})]_{k^\star}| \overset{(61)}{<} 0$$

That being said, $(z_u^\star, z_v^\star)$ is a strict saddle point of Problem (NC$_4$).

Case 2: $\lambda_N^\star \neq 0$. Then by (57), we must have the decoupling property

$$z_u^\star \odot z_v^\star = 0.$$

Therefore,

$$\widetilde{x}_k = \begin{cases} {[z_u^\star]_k}^2 & \text{if } [z_u^\star]_k \neq 0 \Rightarrow [z_v^\star]_k = 0 \\ -{[z_v^\star]_k}^2 & \text{if } [z_v^\star]_k \neq 0 \Rightarrow [z_u^\star]_k = 0 \end{cases}$$

We claim that there must be some $k^\star$ such that

$$[z_u^\star]_{k^\star} = [z_v^\star]_{k^\star} = 0,$$

because otherwise $\widetilde{x}$ has no zero entries, and the equations (60a)–(60b) then implies that

$$\nabla f(\widetilde{x}) = \lambda_N^\star \operatorname{sign}(\widetilde{x})$$

which contradicts with the fact (59):

$$\nabla f(\widetilde{x}) \notin \lambda_N^\star \operatorname{sign}(\widetilde{x})$$

Thus, we not only have $[z_u^\star]_{k^\star} = [z_v^\star]_{k^\star} = 0$ for some $k^\star$, but also

$$|[\nabla f(\widetilde{x})]_{k^\star}| > -\lambda_N^\star > 0, \tag{62}$$

since otherwise we still have $\nabla f(\widetilde{x}) \in \lambda_N^\star \operatorname{sign}(\widetilde{x})$. Then we can choose the direction vector as

$$(d_u, d_v) = \begin{cases} (e_k^\star, 0) & \text{if } [\nabla f(\widetilde{x})]_{k^\star} < 0 \\ (0, e_k^\star) & \text{if } [\nabla f(\widetilde{x})]_{k^\star} > 0 \end{cases}$$

so that we have

$$\begin{bmatrix} d_u^\top & d_v^\top \end{bmatrix} \nabla^2 \mathcal{L}_N(z_u^\star, z_v^\star, \lambda_N^\star) \begin{bmatrix} d_u \\ d_v \end{bmatrix} = -2\,|[\nabla f(\widetilde{x})]_{k^\star}| - 2\lambda_N^\star \overset{(62)}{<} 0$$

That being said, $(z_u^\star, z_v^\star)$ is a strict saddle point of Problem (NC$_4$).

We now complete the proof of Theorem 9. ■

F.4 Extension to nonconvex functions

We can extend our Hadamard parameterization framework to nonconvex objective functions f in Problems (C_1)–(C_4). We make the following assumption.

(P_1) The Hessian $\nabla^2 f(x^\star)$ of Problems (C_1)–(C_4) evaluated at the KKT points $x^\star$ is PSD;

Recall $h(z)$ denotes the Hadamard parametrization given by

$$h(z) = \begin{cases} z \odot z & \text{for } (C_1)\text{–}(C_3) \\ z_u \odot z_u - z_v \odot z_v, & \text{for } (C_4) \end{cases}$$

Theorem 10. *Under Assumption (P_1), any stationary point $z^\star$ of Problems (NC_1)–(NC_4) is a second-order stationary point if and only if $h(z^\star)$ is a KKT point of Problems (C_1)–(C_4).*

Proof. Since the proof of Theorem 10 is quite similar to that of Theorems 1, 7, 8 and 9, to avoid duplication, we will point out which places needed to change in order to adapt the original proofs of Theorems 1, 7, 8 and 9 to the one of Theorem 10.

Note that we will adjust the proof of Theorems 1, 7, 8 and 9 to prove Theorem 10. We aim to establish the one-to-one correspondence between the KKT points of Problems (C_1)–(C_4) and the second-order stationary points of Problems (NC_1)–(NC_4) without making the convexity assumption of f. For this purpose and to write a concise proof, we only make the following changes in the proof of Theorems 1, 7, 8 and 9 (everything else remains unchanged):

1. Let $x^\star$ denote any KKT point of Problems (C_1)–(C_4) instead of their global minimizer. Therefore, we will not use the notations $x^\star \in \text{argmin}_x f(x)$ or $z^\star \in \text{argmin}_z g(z)$ in the main context or the proofs.

2. Denote
$$\mathcal{Z}^\star = \begin{cases} \{z^\star : z^\star \odot z^\star = x^\star\} & \text{for } (C_1)\text{–}(C_3) \\ \{(z_u^\star, z_v^\star) : z_u^\star \odot z_u^\star - z_v^\star \odot z_v^\star = x^\star\} & \text{for } (C_4) \end{cases}$$
where $x^\star$ is any KKT point of Problems (C_1)–(C_4) (instead of a global minimizer).

3. Show (16), (38), (47) and (56) using Assumption (P_1) instead of using the convexity assumption of f.

■

G Detailed Experimental Settings

G.1 Computational Resources

Our large-scale experiments (results presented in Figures 2, 4 and 7) were run on a workstation with an Intel Core i9-9940X CPU and 128GB of RAM. Our small-scale experiments (results presented in Figures 5 and 6) were run on a 2012 MacBook Pro with an Intel Core i5 CPU and 16 GB of RAM. We estimate that approximately 150 hours of CPU time were used to run all our experiments as well as for prototyping. We did not use any GPU resources.

G.2 Implementation of Algorithms

We implemented our proposed algorithms, namely HadRGD, HadRGD-AW and HadRGD-BB, in Python. We used the implementation of Pairwise Frank-Wolfe included in the `copt` package Fabian Pedregosa (2020). We implemented Projected Gradient Descent ourselves and used the algorithm described in (Duchi et al., 2008) to compute projections to Δ_n (see Appendix H.1). We tested two line search schemes: Armijo rule along the feasible direction and Armijo rule along the projection arc (Bertsekas, 1997, Chpt. 3). We found

Algorithm	Hyperparameters	Case (i)	Case (ii)
PGD with linesearch	s	20/L	20/L
	β	0.75	0.75
	ρ_1	10^{-4}	10^{-4}
HadRGD-AW	α_{def}	$10\sqrt{\frac{20n}{L}}$	$10\sqrt{\frac{2n}{L}}$
	β	0.75	0.75
	ρ_1	10^{-4}	10^{-4}
	ρ_2	0.9	0.9
HadRGD-BB	α_{def}	3.0	$10\sqrt{\frac{2n}{L}}$
	β	0.5	0.75
	η	0.5	0.5
	ρ_1	0.1	0.1

Table 1: Hyperparameters for benchmarking described in Section 8. Case (i) is where $x_{\text{true}} \in \text{int}(\Delta_n)$ (see Figure 2). Case (ii) is where x_{true} is on the boundary of Δ_n (see Figure 4). L denotes the Lipschitz constant of the objective function $f(x)$.

the performance of both schemes to be very similar, so in our experiments we used the feasible direction scheme. We implemented Entropic Mirror Descent (EMDA) exactly as described in Section 5 of (Beck and Teboulle, 2003), except we used a constant step-size rule instead of decaying step-sizes.

G.3 Hyperparameters

For EMDA the only hyperparameter is the step size. As mentioned above, we chose to use a constant step size as empirically we observed this led to faster convergence than a decaying step size rule. For the implementation of PFW we used there are no free hyperparameters. For the experiment where x_{true} is on the boundary of Δ_n (see Figure 2) we gave PFW the exact Lipschitz constant of $f(x)$, in an attempt to speed it up (we did not do so when $x_{\text{true}} \in \text{int}(\Delta_n)$.

As PGD with linesearch and HadRGD-AW employ a backtracking line search, there are several important parameters to set. Arguably the most important is the default step size (α_{def} in Algorithms 4, 5 and 6). We chose α_{def} such that the line search sub-routine within PGD and HadRGD-AW took approximately ten iterations to find a step size α_k satisfying the Armijo condition (Armijo-Wolfe conditions for HadRGD-AW). The other hyperparameters were set as described in Table 1. For HadRGD-BB we experimented lightly with hyperparameter tuning, but found that it had little effect on the convergence speed. Thus, we stuck with the somewhat arbitrary values presented in Table 1.

H Additional Experiments

H.1 Projections to the simplex

As discussed in Section 1, the convergence speed of PGD on Δ_n is strongly influenced by the computational complexity of the algorithm used to compute P_{Δ_n}. For completeness, we test four different projection algorithms. Specifically, we test namely the well-known “sort-then-project” algorithm (see Figure 1 in (Duchi et al., 2008), henceforth referred to as `SortProject`)as well as those proposed in (Shalev-Shwartz and Singer, 2006), (Duchi et al., 2008) and (Condat, 2016), henceforth: `PivotProject`, `DuchiProject` and `CondatProject` respectively. All four algorithms are implemented purely in Python. `DuchiProject` and `CondatProject` are exact implementations of the pseudocode presented in (Duchi et al., 2008) and

(Condat, 2016) respectively; `SortProject` is based on the Matlab code provided in (Chen and Ye, 2011) while `PivotProject` is the algorithm `projection_simplex_pivot` available at https://gist.github.com/mblondel/6f3b7aaad90606b98f71. We tested Projected Gradient Descent (PGD) for (C_1) using all four algorithms for the underdetermined least squares problem, as described in Section 8, for true solution x_{true} satisfying case (i) and case (ii). We note that although (Condat, 2016) provides convincing evidence that `CondatProject` is faster than `DuchiProject` when optimized and written in C, our experimental results suggest that `DuchiProject` is the fastest projection algorithm for $x_{\text{true}} \in \text{int}(\Delta)$ while `SortProject` is the fastest for x_{true} on the boundary, when implemented purely in Python. As $x_{\text{true}} \in \text{int}(\Delta)$ is the case of primary interest for us, we use PGD with `DuchiProject` in all our benchmarking experiments.

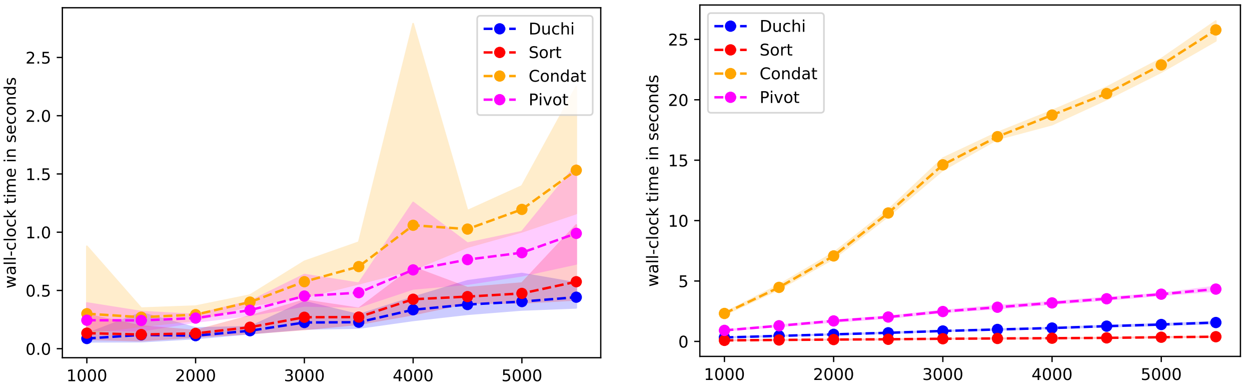

Figure 5: Time required to reach an error less than for PGD with four different projection algorithms. **Left:** $x_{\text{true}} \in \text{int}(\Delta_n)$. We use a target solution accuracy of 10^{-16}. **Right:** x_{true} a projection of a random Gaussian vector to Δ_n. We use a target solution accuracy of 10^{-8} and a maximum of 200 iterations. All results are averaged over ten trials and the shading denotes the min-max range.

H.2 Additional Benchmarking

Figure 6 contains additional benchmarking results where we compare HadRGD and PGD (both without line search) to EMDA and PFW using the so-called Demyanov-Rubinov step-size rule. As EMDA is significantly slower than the other three algorithms, we do not include it in our large-scale benchmarking. Finally, Figure 8 shows HadRGD-AW achieves a linear rate of convergence.

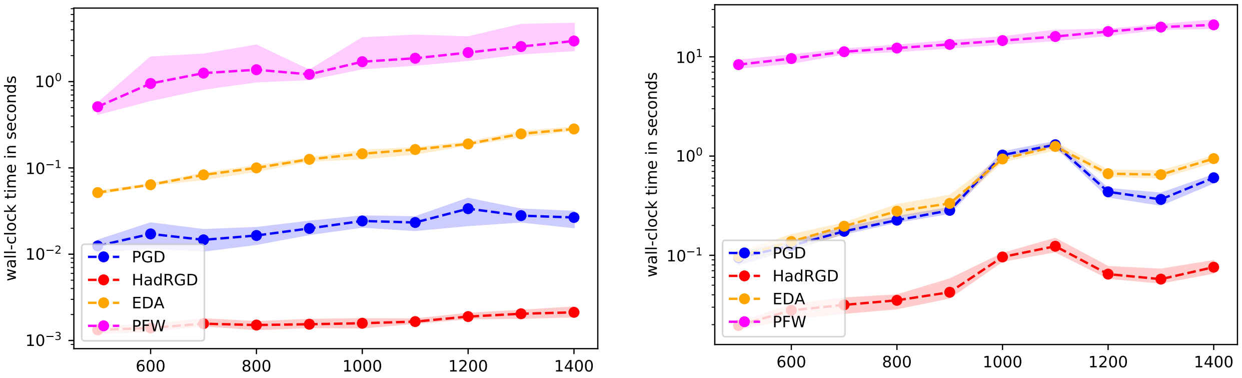

Figure 6: Time required to reach an error less than 10^{-8}, or until the maximum number of iterations is hit. **Left:** $x_{\text{true}} \in \text{int}(\Delta_n)$. **Right:** x_{true} a projection of a random Gaussian vector to Δ_n. All results are averaged over ten trials and the shading denotes the min-max range.

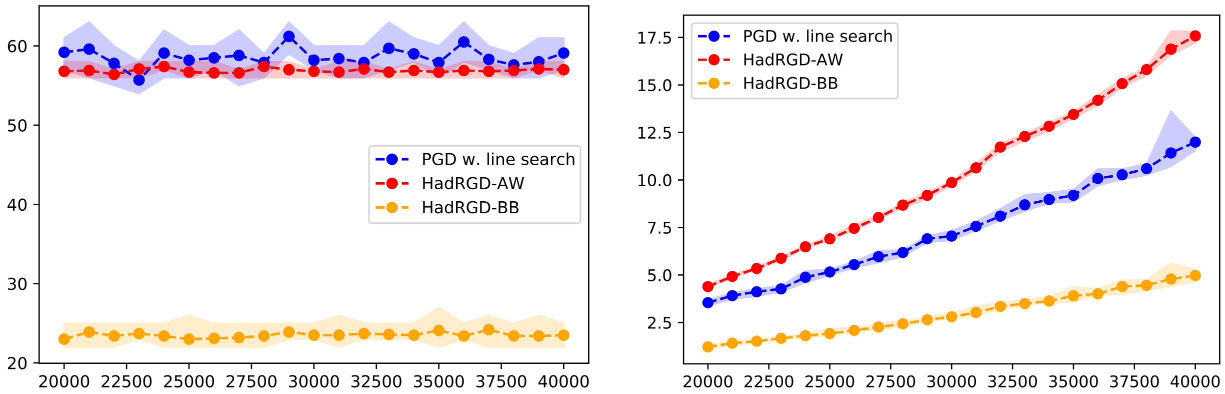

Figure 7: Solving the underdetermined least squares problem as in Section 8 with $x_{\text{true}} \in \text{int}(\Delta)$. Here, the target solution accuracy is 10^{-16}. **Left:** Number of iterations vs. n. **Right:** Wall-clock time vs. n.

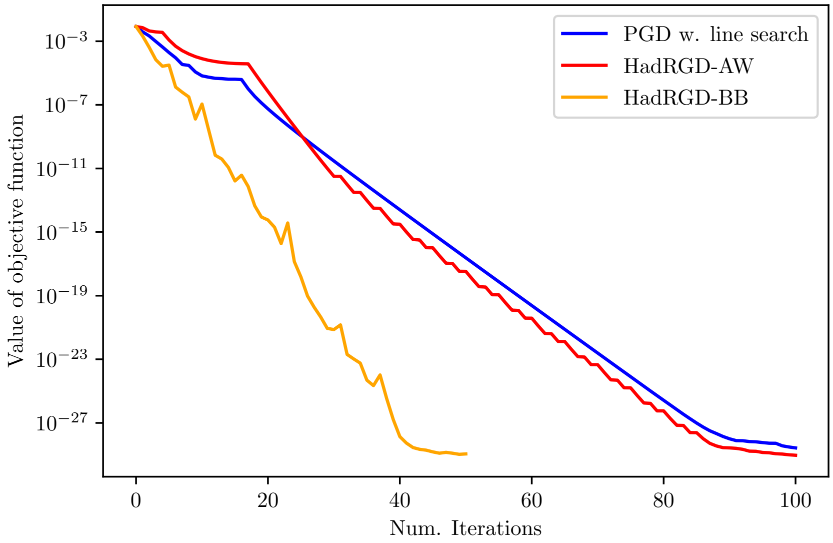

Figure 8: Solving the underdetermined least squares problem as in Section 8 with $x_{\text{true}} \in \text{int}(\Delta)$. HadRGD-AW, as well as HadRGD-BB and PGD with line search, enjoy a linear rate of convergence.