Online algorithm design via smoothing with application to online experiment selection

Reza Eghbali

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Maryam Fazel, Chair

Wyeth Bair

Jeffery A. Bilmes

Mehran Mesbahi

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Online algorithm design via smoothing with application to online experiment selection

Reza Eghbali

Chair of the Supervisory Committee:
Associate Professor Maryam Fazel
Electrical Engineering

In this dissertation, we present the results of our research on three topics, namely, the design and analysis of online convex optimization algorithms, convergence rate analysis of proximal gradient homotopy algorithm for structured convex problems, and application of computational methods for study of brain cells in the visual cortex of the primate brain. In our work on online optimization, we have developed a systematic approach with a clear connection to regret minimization for design and worst case analysis of online optimization algorithms. We apply this approach to online experiment design problems. Our results on the convergence rate analysis of proximal gradient homotopy algorithm extends the linear convergence rate of this algorithm from l_1 norm to a general class of norms called decomposable norms. Our results on the clustering of cells in the visual cortical area V4 reveal new categories of neurons in this area. We discuss how online adaptive algorithms can be utilized for classification of these neurons in closed loop neurophysiological experiments.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

First order methods in optimization – algorithms that only use function value and its derivative – are methods of choice for large scale problems since they have a low per iteration computation cost. While the general convergence rate for first order methods applied to convex functions is sub-linear, in practice, and sometimes in theory, it has been shown that for many structured objective functions arising from problems in machine learning these algorithms demonstrate a local linear convergence rate [DL16, ZS17] and with the right modifications even global linear rates [XZ13].

First order optimization algorithms have also been successfully applied to online optimization and online learning problems. These are optimization problems in which the objective function is modified and new variables are presented online. An online algorithm should assign values to the new variables in a sequential order without the chance to change the previous assignments. The first order algorithm approach to solving online problems gives simple online algorithms that can be analyzed by the same techniques used in the convergence analysis of the first order methods.

The next two chapters of this dissertation concern the two topics mentioned above. In Chapter 2, we study the convergence rate of a variant of the proximal-gradient algorithm applied to norm-regularized linear least squares problems, for a general class of norms. These problems arises in statistical learning for recovery of structured models from small number of linear noisy measurements. We show that if the linear sampling matrix satisfies certain assumptions and the regularizing norm is decomposable, proximal-gradient homotopy algorithm converges with a *linear rate* even though the objective function is not strongly convex. Our result generalizes the result of [XZ13] on the linear convergence of homotopy algorithm

for l_1-regularized least squares problems. Numerical experiments are presented that support the theoretical convergence rate analysis.

In Chapter 3, we present our analysis of the worst case competitive ratio of two primal-dual algorithms for a general class of online convex (conic) optimization problems. Our approach clearly links competitive ratio analysis of online optimization with regret analysis in online learning. We provide new examples of online problems on the positive orthant and the positive semidefinite cone (PSD cone) for which our analysis applies, this includes an important statistical problem, online optimal experiment design. We show how smoothing, which is used to improve the convergence rate of first order algorithms, can improve the competitive ratio of greedy online algorithms. In particular, for separable functions on the positive orthant and trace functions on the PSD cone, we show that the optimal smoothing can be derived by solving a convex optimization problem. This result allows us to directly optimize the competitive ratio bound over a class of smoothing functions, and hence *design* effective smoothing customized for a given cost function.

The last chapter of this dissertation captures our recent work on clustering and characterization of cells in the visual cortical area V4. The cortical area V4 is an intermediate visual area in the ventral (form processing) pathway of the primate visual system. This area is known to be involved in object recognition but is not as well-understood as the earlier stages of the ventral pathway, such as the primary visual cortex (V1). In many studies attempting to understand V4, the stimulus design and analysis were driven by a specific underlying hypothesis and model. To avoid inherent pitfalls in model-based approaches for characterization of V4 neurons, and discover novel encoding dimensions, we use a more model-free approach based on clustering of the cells to reanalyze responses of V4 neurons to a set of simple shapes. This chapter contains the results of our clustering analysis which revealed novel categories of V4 cells. We also study online stimulus sampling algorithms for neurophysiological recordings in V4, using our tools from online optimization and results on clustering of V4 cells.

Chapter 2

DECOMPOSABLE NORM MINIMIZATION

In signal processing and statistical regression, problems arise in which the goal is to recover a structured model from a few, often noisy, linear measurements. Well studied examples include recovery of sparse vectors and low rank matrices. These problems can be formulated as non-convex optimization programs, which are computationally intractable in general. One can relax these non-convex problems using appropriate convex penalty functions, for example ℓ_1, $\ell_{1,2}$ and nuclear norms in sparse vector, group sparse and low rank matrix recovery problems. These relaxations perform very well in many practical applications. Following [Don06, CT06, CRT06], there has been a flurry of publications that formalize the condition for recovery of sparse vectors, e.g., [BTW$^+$07, VDGB$^+$09], low rank matrices, e.g., [RFP10, CP11, Gro11] from linear measurements by solving the appropriate relaxed convex optimization problems. Alongside results for sparse vector and low rank matrix recovery several authors have proposed more general frameworks for structured model recovery problems with linear measurements [CR12, CRPW12, NRWY12]. In many problems of interest, to recover the model from linear noisy measurements, one can formulate the following optimization program:

$$\begin{aligned} \text{minimize} \quad & \|x\| \\ \text{subject to} \quad & \|Ax - b\|_2^2 \leq \epsilon^2, \end{aligned} \tag{2.1}$$

where $b \in \mathbb{R}^m$ is the measurements vector, $A \in \mathbb{R}^{m \times n}$ is the linear measurement matrix, ϵ^2 is the noise energy and $\|\cdot\|$ is a norm on $\mathbb{R}^n$ that promotes the desired structure in the solution. The regularized version of problem (2.1) has the following form:

$$\text{minimize} \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|, \tag{2.2}$$

where $\lambda > 0$ is the regularization parameter.

There has been extensive work on algorithms for solving problem (2.1) and (2.2) in special cases of ℓ_1 and nuclear norms. First order methods have been the method of choice for large scale problems, since each iteration is computationally cheap. Of particular interest is the proximal-gradient method for minimization of composite functions, which are functions that can be written as sum of a differentiable convex function and a closed convex function.

When the smooth component of the objective function has a Lipschitz continuous gradient, proximal-gradient algorithm has a convergence rate of $O(1/t)$, where t is the iteration number. For the accelerated version of proximal-gradient algorithm, the convergence rate improves to $O(1/t^2)$ [Nes13]. When the objective function is strongly convex as well, proximal-gradient has linear convergence, i.e. $O(\kappa^t)$ with $\kappa \in (0, 1)$ [Nes13]. However, in instances of problem (2.2) that are of interest, the number of samples m is less than the dimension of the space n, hence the matrix A has a non-zero null space which results in an objective function that is not strongly convex. Several algorithms that combine homotopy continuation over λ with proximal-gradient steps have been proposed in the literature for problem (2.2) in the special cases of ℓ_1 and nuclear norms [HYZ08, WNF09, WYGZ10, MGC11, TY10]. Xiao and Zhang [XZ13] have studied an algorithm with homotopy with respect to λ for solving ℓ_1 regularized least squares problem. Formulating their algorithm based on Nesterov's proximal-gradient method, they have demonstrated that this algorithm has an overall linear rate of convergence whenever A satisfies the restricted isometry property (RIP) and the final value of the regularizer parameter λ is greater than a problem-dependent lower bound.

2.0.1 Our result

We generalize the linear convergence rate analysis of the homotopy algorithm studied in [XZ13] to problem (2.2) when the regularizing norm is decomposable, where decomposability is a condition introduced in [CR12]. In particular, ℓ_1, $\ell_{1,2}$ and nuclear norms satisfy this condition. We derive properties for this class of norms that are used directly in the convergence analysis. These properties can independently be of interest. Among these properties

is the sublinearity of the the function $K : \mathbb{R}^n \mapsto \{0, 1, \ldots, n\}$, where K is generalization of the notion of cardinality for decomposable norms defined in (2.12).

The linear convergence result holds under an assumption on the RIP constants of A, which in turn holds with high probability for several classes of random matrices when the number of measurements m is large enough (orderwise the same as that required for recovery of the structured model).

2.0.2 Algorithms for structured model recovery

There has been extensive work on algorithms for solving problems (2.1) and (2.2) in the special cases of ℓ_1 and nuclear norms. For a detailed review of first order methods we refer the reader to [NN13] and references therein. In [XZ13], authors have reviewed sparse recovery and ℓ_1 norm minimization algorithms that are related to the homotopy algorithm for ℓ_1 norm. We discuss related algorithms mostly focusing on algorithms for other norms including nuclear norm here.

The proximal-gradient method for ℓ_1/nuclear norm minimization has a local linear convergence in a neighborhood of the optimal value [HZSL13, ZJL13, LT92]. The proximal operator for nuclear norm is soft-thresholding operator on singular values. Several authors have proposed algorithms for low rank matrix recovery and matrix completion problem based on soft- or hard-thresholding operators; see, e.g., [JMD10, CCS10, MHT10, MGC11]. The singular value projection algorithm proposed by Jain et al. has a linear rate; however, to apply the hard-thresholding operator, one should know the rank of x_0. While the authors have introduced a heuristic for estimating the rank when it is not known a priori, their convergence results rely upon a known rank [JMD10]. SVP is the generalization of iterative hard thresholding algorithm (IHT) for sparse vector recovery. SVP and IHT belong to the family of greedy algorithms which do not solve a convex relaxation problem. Other greedy algorithms proposed for sparse recovery such as Compressive Sampling Matching Pursuit (CoSaMP) [NT09] and Fully Corrective Forward Greedy Selection (FCFGS) [SSSZ10] have also been generalized for recovery of general structured models including low-rank matrices

and extended to more general loss functions [NNW14, NHNT13, SSGS11].

For huge-scale problems with separable regularizing norm such as ℓ_1 and $\ell_{1,2}$, coordinate descent methods can reduce the computational cost of each iteration significantly. The convergence rate of randomized proximal coordinate descent method in expectation is orderwise the same as full proximal gradient descent; however, it can yield an improvement in terms of the dependence of convergence rate on n [Nes12, RT14, LX13]. To the best of our knowledge, linear convergence rate for any coordinate descent method applied to problem (2.1) or (2.2) has not been shown in the literature.

Continuation over λ for solving the regularized problem has been utilized in fixed point continuation algorithm (FPC) proposed by Ma et al. [MGC11] and accelerated proximal-gradient algorithm with line search (APGL) by Toh et al. [TY10]. FPC and APGL both solve a series of regularized problems where in each outer-iteration λ is reduced by a factor less than one, the former uses soft-thresholding and the latter uses accelerated proximal-gradient for solving each regularized problem.

Agarwal et al. [ANW10] have proposed algorithms for solving problems (2.1) and (2.2) with an extra constraint in the form of $\|x\| \leq \rho$. They have introduced the assumption of decomposability of the norm and give convergence analysis for norms that satisfy that assumption. They establish linear rate of convergence for their algorithms up to a neighborhood of the optimal solutions. However, their algorithm uses the bound ρ which should be selected based on the norm of the true solution. In many problems this quantity is not known beforehand. Jin et al. [JYZ13] have proposed an algorithm for ℓ_1 regularized least squares that receives ρ as a parameter and has linear rate of convergence. Their algorithm utilizes proximal gradient method but unlike homotopy algorithm reduces λ at each step.

By using the SDP formulation of nuclear norm, interior point methods can be utilized to solve problems (2.1) and (2.2). Interior point methods do not scale as well as first order methods (for example, for a general SDP solver when the dimension exceeds a few hundred). However, Specialized SDP solvers for nuclear norm minimization can bring down the computational complexity of each iteration to $O(n^3)$ [LV09].

2.1 Preliminaries

Let $A \in \mathbb{R}^{m\times n}$. We equip $\mathbb{R}^n$ with an inner product given by $\langle x, y\rangle = x^T By$ for some positive definite matrix B. We equip $\mathbb{R}^m$ with ordinary dot product $\langle v, u\rangle = v^T u$. We denote the adjoint of A as $A^* = B^{-1}A^T$. Note that for all $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$

$$\langle Ax, u\rangle = \langle x, A^* u\rangle. \tag{2.3}$$

We use $\|\cdot\|_2$ to denote the norms induced by the inner product in $\mathbb{R}^n$ and $\mathbb{R}^m$, that is:

$$\begin{aligned} \forall x \in \mathbb{R}^n : &\quad \|x\|_2 = \sqrt{x^T Bx}, \\ \forall v \in \mathbb{R}^m : &\quad \|v\|_2 = \sqrt{v^T v}. \end{aligned}$$

We use $\|\cdot\|$ and $\|\cdot\|^*$ to denote a regularizing norm and its dual on $\mathbb{R}^n$. The latter is defined as:

$$\|y\|^* = \sup\{\langle y, x\rangle \mid \|x\| \le 1\}.$$

Given a convex function $f : \mathbb{R}^n \mapsto \mathbb{R}$, $\partial f(x)$ denotes the set of subgradients of f at x, i.e., the set of all $z \in \mathbb{R}^n$ such that

$$\forall y \in \mathbb{R}^n : \quad f(y) \ge f(x) + \langle z, y - x\rangle.$$

When f is differentiable, $\partial f(x) = \{\nabla f(x)\}$. Note that $\xi \in \partial\|x\|$ if and only if

$$\langle \xi, x\rangle = \|x\|, \tag{2.4}$$

$$\|\xi\|^* \le 1. \tag{2.5}$$

We say f is strongly convex with strong convexity parameter μ_f when $f(x) - \frac{\mu_f}{2}\|x\|_2^2$ is convex. For a differentiable function this implies that for all $x, y \in \mathbb{R}^n$:

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu_f}{2}\|x - y\|_2^2. \tag{2.6}$$

We call the gradient of a differentiable function Lipschitz continuous with Lipschitz constant L_f, when for all $x, y \in \mathbb{R}^n$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L_f \|y - x\|_2. \tag{2.7}$$

For a convex function f, gradient Lipschitz continuity is equivalent to the following inequality [see [NN04] Lemma 1.2.3. and Theorem 2.1.5]:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f}{2}\|x - y\|_2^2, \tag{2.8}$$

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

2.2 Properties of the regularizing norm and A

In this section we introduce our assumptions on the regularizing norm $\|\cdot\|$, and derive the properties of the norm based on these assumptions. The homotopy algorithm of [XZ13] for the ℓ_1-regularized problem is designed so that the iterates maintain low cardinality throughout the algorithm, therefore one can use the restricted eigenvalue property of A, when A acts on these iterates. Said another way, the squared loss term behaves like a strongly convex function over the algorithm iterates, which is why the algorithm can achieve a fast convergence rate. In the proof, [XZ13] uses the structure of the subdifferential of the ℓ_1 norm,

$$\partial\|x\|_1 = \{\text{sgn}(x) + v \mid v_i = 0 \text{ when } x_i \neq 0,\ \|v\|_\infty \leq 1\},$$

as well as the following properties that hold for the cardinality function,

$$\|x\|_1^2 \leq \text{card}(x)\|x\|_2^2,$$
$$\text{card}(x + y) \leq \text{card}(x) + \text{card}(y) \quad \text{(sublinearity)}.$$

We first give our assumption on the structure of the subdifferential of a class of norms (which inlcudes ℓ_1 and nuclear norms but is much more general), and then derive the rest of the properties needed for generalizing the results of [XZ13].

Before stating our assumptions, we add some more definitions to our tool box. Let $S^{n-1} = \{x \in \mathbb{R}^n \,|\, \|x\|_2 = 1\}$, and let $\mathcal{G}_{\|\cdot\|}$ be the set of extreme points of the norm ball $\mathcal{B}_{\|\cdot\|} := \{x \mid \|x\| \leq 1\}$. We impose two conditions on the regularizing norm.

Condition 1. *For any $x \in \mathcal{G}_{\|\cdot\|}$, $\|x\|_2 = 1$, i.e., all the extreme points of the norm ball have unit $\|\cdot\|_2$-norm.*

The second condition on the norm is the decomposability condition introduced in [CR12], which was inspired by the assumption introduced in [NRWY12].

Condition 2 (Decomposability)**.** *For all $x \in \mathbb{R}^n$, there exists a subspace T_x and a vector $e_x \in T_x$ such that*

$$\partial\|x\| = \{e_x + v \mid v \in T_x^{\perp}, \|v\|^* \leq 1\}. \tag{2.9}$$

Note that $x \in T_x$ for all $x \in \mathbb{R}^n$ because if $x \notin T_x$, then $x = y + z$ with $y \in T_x$ and $z \in T_x^{\perp} - \{0\}$. Let $z' = z/\|z\|^*$. Since $e_x + z' \in \partial\|x\|$, $\|x\| = \langle e_x + z', y + z\rangle = \|x\| + \|z\|_2^2/\|z\|^*$, which is a contradiction.

The decomposability condition has been used in both [CR12] and [NRWY12] to give a simpler and unified proof for recovery of several structures such as sparse vectors and low-rank matrices. The term decompsable refers to the fact that subgradients can be decomposed into components given in (2.9).

When attempting to extend this algorithm to general norms, several challenges arise. First, what is the appropriate generalization of cardinality for other structures and their corresponding norms? Essentially, we would need to count the number of nonzero coefficients in an appropriate representation and ensure there is a small number of nonzero coefficients in our iterates, to be able to apply a similar proof idea as in [XZ13].

The next theorem captures one of our main results for any decomposable norm. This theorem provides a new set of conditions that is based on the geometry of the norm ball, and we show are equivalent to decomposability on $\mathbb{R}^n$. As a result, one can find a decomposition for any vector in $\mathbb{R}^n$ in terms of an orthogonal subset of $\mathcal{G}_{\|\cdot\|}$.

Theorem 1 (Orthogonal representation)**.** *Suppose $\mathcal{G}_{\|\cdot\|} \subset S^{n-1}$, then $\|\cdot\|$ is decomposable if and only if for any $x \in \mathbb{R}^n - \{0\}$ and $a_1 \in \operatorname{argmax}_{a \in \mathcal{G}_{\|\cdot\|}} \langle a, x\rangle$ there exist $a_2, \ldots, a_k \in \mathcal{G}_{\|\cdot\|}$ such that $\{a_1, a_2, \ldots, a_k\}$ is an orthogonal set that satisfies the following conditions:*

I There exists $\{\gamma_i > 0 | \, i = 1, \ldots, k\}$ *such that:*

$$x = \sum_{i=1}^{k} \gamma_i a_i,$$
$$\|x\| = \sum_{i=1}^{k} \gamma_i. \tag{2.10}$$

II For any set $\{\eta_i \, | \, |\eta_i| \leq 1, i = 1, \ldots, k\}$*:*

$$\left\| \sum_{i=1}^{k} \eta_i a_i \right\|^* \leq 1. \tag{2.11}$$

Moreover, if $\{a_1, a_2, \ldots, a_k\} \subset \mathcal{G}_{\|\cdot\|}$ *satisfy I and II, then* $e_x = \sum_{i=1}^{k} a_i$*.*

The proof of Theorem 1 is presented in Appendix B.

We will see in section 2.4 that we need an orthogonal representation for all vectors to be able to bound the number of nonzero coefficients throughout the algorithm. First, we define a quantity $K(x)$ that bounds the ratio of the norm $\|\cdot\|$ to the Euclidean norm, and plays the same role in our analysis as cardinality played in [XZ13]. Then we show that $K(x)$ is a sublinear function, that is, $K(x+y) \leq K(x) + K(y)$ for all x, y. This is a key property that is needed in the convergence analysis. Define $K : \mathbb{R}^n \mapsto \{0, 1, 2, \ldots, n\}$

$$K(x) = \|e_x\|_2^2. \tag{2.12}$$

Note that for every $x \in \mathbb{R}^n$,

$$\|x\|^2 = \langle e_x, x \rangle^2 \leq \|e_x\|_2^2 \|x\|_2^2 = K(x)\|x\|_2^2. \tag{2.13}$$

Here, the first equality follows from (2.4), and the inequality follows from the Cauchy-Schwarz inequality. In the analysis of homotopy algorithm we utilize (2.13) alongside the structure of the subgradient given by (2.9). ℓ_1, $\ell_{1,2}$, and nuclear norms are three important examples that satisfy conditions 1 and 2. Here we briefly discuss each one of these norms.

- **Nuclear norm** on $\mathbb{R}^{d_1 \times d_2}$ is defined as

$$\|X\|_* = \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i(X)$$

Where $\sigma_i(X)$ is the i^{th} largest singular value of X given by the singular value decomposition $X = \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i(X)\, u_i v_i^T$. With the trace inner product $\langle X, Y\rangle = \text{trace}\left(X^T Y\right)$, nuclear norm satisfies conditions 1 and 2. In this case, $K(X) = \text{rank}(X)$, $\gamma_i = \sigma_i(X)$ and $a_i = u_i v_i^T$ for $i \in \{1, 2, \ldots, \text{rank}(X)\}$. The subspace T_X is given by:

$$T_X = \left\{ \sum_{i=1}^{\text{rank}(X)} u_i z_i^T + z'_i v_i^T \;\middle|\; z_i \in \mathbb{R}^{d_2}, z'_i \in \mathbb{R}^{d_1}, \text{ for all } i \right\},$$

while $e_X = \sum_{i=1}^{\text{rank}(X)} u_i v_i^T$.

- **Weighted ℓ_1 norm** on $\mathbb{R}^n$ is defined as:

$$\|x\|_1 = \sum_{i=1}^{n} w_i |x_i|$$

where w is a vector of positive weights. With $\langle x, y\rangle = \sum_{i=1}^n w_i^2 x_i y_i$, ℓ_1 norm satisfies conditions 1 and 2. For ℓ_1 norm, $K(x) = |\{i | x_i \neq 0\}|$, $\{\gamma_1, \gamma_2, \ldots, \gamma_k\} = \{w_i |x_i| \mid |x_i| > 0, i = 1, \ldots, n\}$. T_x is the support of x, which is defined as:

$$T_x = \{y \in \mathbb{R}^n \mid y_i = 0 \text{ if } x_i = 0\},$$

while the i^{th} element of e_x is $\text{sign}(x_i) w_i$.

- **$\ell_{1,2}$ norm on $\mathbb{R}^{d_1 \times d_2}$:** For a given inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \mapsto \mathbb{R}$ and its induced norm $\|\cdot\|_2$ on $\mathbb{R}^{d_1}$, We define:

$$\|X\|_{1,2} = \sum_{i=1}^{d_2} \|X_i\|_2,$$

where X_i denotes the i^{th} column of X. With inner product $\langle X, Y\rangle = \sum_{i=1}^{d_2} \langle X_i, Y_i\rangle$, $\ell_{1,2}$ norm satisfies conditions 1 and 2. For this norm, $K(X) = |\{i | X_i \neq 0\}|$ and

$\{\gamma_1, \gamma_2, \ldots, \gamma_k\} = \{\|X_i\|_2 \mid \|X_i\|_2 > 0, i = 1, \ldots, d_2\}$. T_X is the column support of X, which is defined as:

$$T_X = \left\{ [Y_1, Y_2, \ldots, Y_{d_2}] \in \mathbb{R}^{d_1 \times d_2} \;\middle|\; Y_i = 0 \text{ if } X_i = 0 \right\},$$

while the i^{th} column of e_X is equal to 0 if $X_i = 0$ and is equal to $X_i / \|X_i\|_2$ otherwise.

Our second result on properties of decomposable norms is captured in the next theorem which establishes sublinearity of K for decomposable norms.

Theorem 2. *For all $x, y \in \mathbb{R}^n$*

$$K(x + y) \leq K(x) + K(y). \tag{2.14}$$

Theorem 2 for ℓ_1, $\ell_{1,2}$ and nuclear norm is equivalent to sublinearity of cardinality of vectors, number of non-zero columns and rank of matrices. The proof of this theorem is included in Appendix B.

2.2.1 Properties of A

Restricted Isometry Property was first discussed in [CT06] for sparse vectors. Generalization of that concept to low rank matrices was introduced in [RFP10]. Note that if $K(x) \leq k$, then $\|x\| \leq \sqrt{k}\|x\|_2$. Based on this observation we define restricted isometry constants of $A \in \mathbb{R}^{m \times n}$ as:

Definition 1. *The upper (lower) restricted isometry constant $\rho_+(A, k)$ ($\rho_-(A, k)$) of a matrix $A \in \mathbb{R}^{m \times n}$ is the smallest (largest) positive constant that satisfies this inequality:*

$$\rho_-(A, k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq \rho_+(A, k)\|x\|_2^2,$$

whenever $\|x\|^2 \leq k\|x\|_2^2$.

Proposition 1. *Let $A \in \mathbb{R}^{m \times n}$ and $f(x) = \frac{1}{2}\|Ax - b\|_2^2$. Suppose that $\rho_+(A, k)$ and $\rho_-(A, k)$ are restricted isometry constants corresponding to A, then:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\rho_-(A, k)\|x - y\|_2^2, \tag{2.15}$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\rho_+(A,k)\|x-y\|_2^2, \tag{2.16}$$

for all $x, y \in \mathbb{R}^n$ *such that* $\|x-y\|^2 \leq k\|x-y\|_2^2$.

Proposition (1) follows from the definition of restricted isometry constants and the following equality:

$$\frac{1}{2}\|A(x-y)\|_2^2 = f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

2.3 Proximal-gradient method and homotopy algorithm

We state the proximal-gradient method and the homotopy algorithm for the following optimization problem:

$$\text{minimize} \quad \phi_\lambda(x) = f(x) + \lambda\|x\|,$$

where $f(x) = \frac{1}{2}\|Ax-b\|_2^2$. While, for simplicity, we analyze the homotopy algorithm for the least squares loss function, the analysis can be extended to every function of form $f(x) = g(Ax)$ when g is a differentiable strongly convex function with Lipschitz continuous gradient. The key element in the proximal-gradient method is the proximal operator which was developed by Moreau [Mor62] and later extended to maximal monotone operators by Rockafellar [Roc76]. Nesterov has proposed several variants of the proximal-gradient methods [Nes13]. In this section, we discuss the gradient method with adaptive line search. For any $x, y \in \mathbb{R}^n$ and positive L, we define:

$$\begin{aligned} m_{\lambda,L}(y,x) &= f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x-y\|_2^2 + \lambda\|x\|, \\ \text{Prox}_{\lambda,L}(y) &= \operatorname*{argmin}_{x \in \mathbb{R}^n} m_{\lambda,L}(y,x) \\ \omega_\lambda(x) &= \min_{\xi \in \partial\|x\|} \|\lambda\xi + \nabla f(x)\|^*. \end{aligned}$$

Xiao and Zhang [XZ13] have considered the proximal-gradient homotopy algorithm for ℓ_1 norm. Here we state it for general norms. Algorithm (1), introduces the homotopy algorithm and contains the proximal-gradient method as a subroutine. The stopping criteria

in the proximal-gradient method is based on the quantity

$$\left\|M_t(x^{(t-1)} - x^{(t)}) + \nabla f(x^{(t)}) - \nabla f(x^{(t-1)})\right\|^*,$$

which is an upper bound on $\omega_\lambda(x^{(t)})$. This follows from the fact that since $x^{(t)} = \operatorname{argmin}_{x\in\mathbb{R}^n} m_{\lambda,M_t}(x^{(t-1)}, x)$, there exists $\xi \in \partial\left\|x^{(t)}\right\|$ such that $\nabla f(x^{(t-1)}) + \lambda\xi + M_t(x^{(t)} - x^{(t-1)}) = 0$. Therefore,

$$\begin{aligned}\omega_\lambda(x^{(t)}) \leq \left\|\lambda\xi + \nabla f(x^{(t)})\right\|^* &= \left\|\lambda\xi + \nabla f(x^{(t-1)}) + \nabla f(x^{(t)}) - \nabla f(x^{(t-1)})\right\|^* \\ &\leq \left\|M_t(x^{(t-1)} - x^{(t)}) + \nabla f(x^{(t)}) - \nabla f(x^{(t-1)})\right\|^*. \end{aligned} \tag{2.17}$$

The homotopy algorithm reduces the value of λ in a series of steps and in each step applies the proximal-gradient method. At step t, $\lambda_t = \lambda_0\eta^t$ and $\epsilon_t = \delta'\lambda_t$ with $\eta \in (0,1)$ and $\delta' \in (0,1)$. In the proximal-gradient method and the backtracking subroutine, the parameters $\gamma_{\text{dec}} \geq 1$ and $\gamma_{inc} > 1$ should be initialized. Since the function f satisfies the inequality (2.8), it is clear that $L_{\min}$ should be chosen less than L_f.

Theorem 5 in [Nes13] states that the proximal-gradient method has a linear rate of convergence when f satisfies (2.6) and (2.8). In proposition 2 we restate that theorem with minimal assumptions which is f satisfies (2.6) and (2.8) on a restricted set. The proof of this proposition is given in appendix B.

Proposition 2. *Let $x^* \in \operatorname{argmin}\phi_\lambda$. If for every t:*

$$f\left(x^{(t)}\right) \geq f\left(x^*\right) + \left\langle \nabla f\left(x^*\right), x^{(t)} - x^*\right\rangle + \frac{\mu_f}{2}\left\|x^{(t)} - x^*\right\|_2^2, \tag{2.18}$$

$$f\left(x^{(t+1)}\right) \geq f\left(x^{(t)}\right) + \left\langle \nabla f\left(x^{(t)}\right), x^{(t+1)} - x^{(t)}\right\rangle + \frac{\mu_f}{2}\left\|x^{(t)} - x^{(t+1)}\right\|_2^2, \tag{2.19}$$

$$f\left(x^{(t+1)}\right) \leq f\left(x^{(t)}\right) + \left\langle \nabla f\left(x^{(t)}\right), x^{(t+1)} - x^{(t)}\right\rangle + \frac{L_f}{2}\left\|x^{(t)} - x^{(t+1)}\right\|_2^2, \tag{2.20}$$

then

$$\phi_\lambda\left(x^{(t)}\right) - \phi_\lambda\left(x^*\right) \leq \left(1 - \frac{\mu_f\gamma_{\text{inc}}}{4L_f}\right)^t \left(\phi_\lambda\left(x^{(0)}\right) - \phi_\lambda\left(x^*\right)\right). \tag{2.21}$$

In addition, if

$$\left\|\nabla f\left(x^{(t)}\right) - \nabla f\left(x^{(t+1)}\right)\right\|^* \leq L_f'\left\|x^{(t)} - x^{(t+1)}\right\|_2 \tag{2.22}$$

Algorithm 1 Homotopy

Input: $\lambda_{\text{tgt}} > 0$, $\epsilon > 0$

Parameters: $\eta \in (0,1)$, $\delta' \in (0,1)$, $L_{\min} > 0$

$y^{(0)} \leftarrow 0$, $\lambda_0 \leftarrow \|A^* b\|^*$, $M \leftarrow L_{\min}$, $N \leftarrow \lfloor \log\left(\frac{\lambda_{\text{tgt}}}{\lambda_0}\right) / \log(\eta) \rfloor$

for $t = 0, 1, \ldots, N-1$ **do**

$\lambda_{t+1} \leftarrow \eta \lambda_t$

$\epsilon_t \leftarrow \delta' \lambda_t$

$[y^{(t+1)}, M] \leftarrow \text{ProxGrad\_}\phi_{\lambda_{t+1}}\left(y^{(t)}, M, L_{\min}, \epsilon_t\right)$

end for

$[y, M] \leftarrow \text{ProxGrad\_}\phi_{\lambda_{\text{tgt}}}\left(y^{(N)}, M, L_{\min}, \epsilon\right)$

Subroutine 1 $[x, M] = \text{ProxGrad\_}\phi_\lambda\left(x^{(0)}, L_0, L_{\min}, \epsilon'\right)$

Parameter: $\gamma_{\text{dec}} \geq 1$,

$t \leftarrow 0$

repeat

$[x^{(t+1)}, M_{t+1}] \leftarrow \text{Backtrack}_{\text{-}\phi_\lambda}\left(x^{(t)}, L_t\right)$

$L_{t+1} \leftarrow \max\{L_{\min}, M_{t+1}/\gamma_{\text{dec}}\}$

$t \leftarrow t + 1$

until $\left\|M_t(x^{(t-1)} - x^{(t)}) + \nabla f\left(x^{(t)}\right) - \nabla f\left(x^{(t-1)}\right)\right\|^* \leq \epsilon'$

$x \leftarrow x^{(t)}$, $M \leftarrow M_t$

Subroutine 2 $[y, M] = \text{Backtrack\_}\phi_\lambda\ (x, L)$

Parameter: $\gamma_{\text{inc}} > 1$

while $\phi_\lambda\left(\text{Prox}_{\lambda,L}(x)\right) > m_{\lambda,L}\left(x, \text{Prox}_{\lambda,L}(x)\right)$ **do**

$L \leftarrow \gamma_{\text{inc}} L$

end while

$y \leftarrow \text{Prox}_{\lambda,L}(x)$, $M \leftarrow L$

and

$$\left\|x^{(t)} - x^{(t+1)}\right\|^* \leq \theta \left\|x^{(t)} - x^{(t+1)}\right\|_2 \tag{2.23}$$

for some constants θ *and* L'_f*, then*

$$\begin{aligned}\omega_\lambda\left(x^{(t+1)}\right) &\leq \left\|M_{t+1}(x^{(t)} - x^{(t+1)}) + \nabla f\left(x^{(t+1)}\right) - \nabla f\left(x^{(t)}\right)\right\|^* \\ &\leq \theta\left(1 + \frac{L'_f}{\mu_f}\right)\sqrt{2\gamma_{\text{inc}} L_f\left(\phi_\lambda\left(x^{(t)}\right) - \phi_\lambda\left(x^*\right)\right)}.\end{aligned} \tag{2.24}$$

2.4 Convergence result

First note that since the objective function is not strongly convex if one applies the sublinear convergence rate of proximal gradient method, the iteration complexity of the homotopy algorithm is $O(\frac{1}{\epsilon} + \sum_{t=1}^{N} \frac{1}{\delta' \lambda_t})$ which can be simplified to $O(\frac{1}{\epsilon} + \frac{1}{\delta'(1-\eta)\lambda_{tgt}})$. As it was stated in the introduction, we use the structure of this problem to provide a linear rate of convergence when assumptions similar to those needed to derive recovery bounds hold.

Suppose $b = Ax_0 + z$, for some $x_0 \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$. Here, z is the noise vector that is added to linear measurements from an structured model x_0. Also, we define $k_0 := K(x_0)$ and the constant c:

$$c := \max_{x \in T_{x_0} - \{0\}} \frac{\|x\|^2}{k_0 \|x\|_2^2}.$$

Note that $c = 1$ for ℓ_1 and $\ell_{1,2}$ norms, and $c \leq 2$ for nuclear norm. This follows from the fact that $K(x) = k_0$ when $x \in T_{x_0}$ for ℓ_1, $\ell_{1,2}$ norms, while $K(x) \leq 2k_0$ when $x \in T_{x_0}$ in case of nuclear norm. Through out this section, we assume the regularizing norm satisfies conditions 1 and 2 introduced in Section 2.2. Before we state the convergence theorem, we introduce an assumption:

Assumption 1. λ_{tgt} *is such that* $\|A^* z\|^* \leq \frac{\lambda_{\text{tgt}}}{4}$*. Furthermore, there exist constants* $r > 1$ *and* $\delta \in (0,\ \frac{1}{4}]$ *such that:*

$$\frac{\rho_-\left(A, ck_0\left(1+\gamma\right)^2\right)}{\rho_+\left(A, 72rck_0(1+\gamma)\gamma_{\text{inc}}\right)} > \frac{c}{r} \tag{2.25}$$

$$\rho_-\left(A, 72rck_0(1+\gamma)\gamma_{\text{inc}}\right) > 0 \tag{2.26}$$

where:

$$\gamma := \frac{\lambda_{\text{tgt}}(1+\delta) + \|A^*z\|^*}{\lambda_{\text{tgt}}(1-\delta) - \|A^*z\|^*}. \tag{2.27}$$

We define $\tilde{k} = 36rck_0(1+\gamma)\gamma_{\text{inc}}$. In appendix A, we provide an upper bound on the number of measurement needed for (2.25) to be satisfied with high probability whenever rows of A are sub-Gaussian random vectors.

The next theorem establishes the linear convergence of the proximal gradient method when $\omega_\lambda\left(x^{(0)}\right) = \min_{\xi\in\partial\left\|x^{(0)}\right\|} \|\nabla f(x) + \lambda\xi\|^*$ is sufficiently small, while Theorem 4 establishes the overall linear rate of convergence of homotopy algorithm.

Theorem 3. *Let $x^{(t)}$ denote the t^{th} iterate of ProxGrad_ϕ_λ $\left(x^{(0)}, L_0, L_{\min}, \epsilon'\right)$, and let $x^* \in \operatorname{argmin}\phi_\lambda(x)$. Suppose Assumption 1 holds true for some r and δ, $L_{\min} \le \gamma_{\text{inc}}\rho_+\left(A, 2\tilde{k}\right)$, and $\lambda \ge \lambda_{\text{tgt}}$. If $x^{(0)}$ satisfies:*

$$K\left(x^{(0)}\right) \le \tilde{k}, \quad \omega_\lambda\left(x^{(0)}\right) \le \delta\lambda,$$

then:

$$K\left(x^{(t)}\right) \le \tilde{k}, \tag{2.28}$$

$$\phi_\lambda\left(x^{(t)}\right) - \phi_\lambda(x^*) \le \left(1 - \frac{1}{4\gamma_{\text{inc}}\kappa}\right)^t \left(\phi_\lambda\left(x^{(0)}\right) - \phi_\lambda(x^*)\right), \tag{2.29}$$

and

$$\omega_\lambda\left(x^{(t)}\right) \le \left(1 + \frac{\sqrt{\rho_+(A,1)\,\rho_+\left(A,2\tilde{k}\right)}}{\rho_-\left(A,2\tilde{k}\right)}\right)\sqrt{2\gamma_{\text{inc}}\rho_+\left(A,2\tilde{k}\right)\left(\phi_\lambda\left(x^{(t-1)}\right) - \phi_\lambda(x^*)\right)}, \tag{2.30}$$

where $\kappa = \frac{\rho_+\left(A,2\tilde{k}\right)}{\rho_-\left(A,2\tilde{k}\right)}$.

Theorem 4. *Let $y^{(t)}$ denote the t^{th} iterate of Homotopy algorithm, and let $y^* \in \operatorname{argmin}\phi_{\lambda_{\text{tgt}}}(y)$. Suppose Assumption 1 holds true for some r and δ, $L_{\min} \le \gamma_{\text{inc}}\rho_+\left(A, 2\tilde{k}\right)$, and $\lambda_0 \ge \lambda_{tgt}$. Furthermore, suppose that δ' and η in the algorithm satisfy:*

$$\frac{1+\delta'}{1+\delta} \le \eta. \tag{2.31}$$

When $t = 0, 1, \ldots, N-1$, the number of proximal-gradient iterations for computing $y^{(t)}$ is bounded by

$$\frac{\log\left(C/\delta^2\right)}{\log\left(1-\frac{1}{4\gamma_{\text{inc}}\kappa}\right)^{-1}}, \tag{2.32}$$

The number of proximal-gradient iterations for computing y is bounded by

$$\frac{\log\left(C\lambda_{\text{tgt}}/\epsilon^2\right)}{\log\left(1-\frac{1}{4\gamma_{\text{inc}}\kappa}\right)^{-1}}, \tag{2.33}$$

where $C := 6\gamma_{\text{inc}}\kappa\delta ck_0\left(1+\gamma\right)\left(\sqrt{\rho_-\left(A,2\tilde{k}\right)}+\sqrt{\rho_+\left(A,1\right)\kappa}\right)^2\Big/\rho_-\left(A,c\left(1+\gamma\right)^2k_0\right)$ *and* $\kappa=\frac{\rho_+\left(A,2\tilde{k}\right)}{\rho_-\left(A,2\tilde{k}\right)}$. *The objective gap of the output y is bounded by*

$$\phi_{\lambda_{\text{tgt}}}\left(y\right)-\phi_{\lambda_{\text{tgt}}}\left(y^*\right)\leq\frac{9ck_0\lambda_{\text{tgt}}\left(1+\gamma\right)\epsilon}{\rho_-\left(A,c\left(1+\gamma\right)^2k_0\right)},$$

while the total number of iterations for computing y is bounded by:

$$\frac{\log\left(C\lambda_{\text{tgt}}/\epsilon^2\right)+\left(\log\left(\frac{\lambda_{\text{tgt}}}{\lambda_0}\right)/\log\left(\eta\right)\right)\log\left(C/\delta^2\right)}{\log\left(1-\frac{1}{4\gamma_{\text{inc}}\kappa}\right)^{-1}}.$$

2.4.1 Parameters selection satisfying the assumptions

Four parameters of $L_{\min}$, λ_{tgt}, δ' and η should be set in the homotopy algorithm. The assumption on $L_{\min}$ is only for convenience. If $L_{\min} > \gamma_{\text{inc}}\rho_+\left(A,2\tilde{k}\right)$, one can replace $\gamma_{\text{inc}}\rho_+\left(A,2\tilde{k}\right)$ with $L_{\min}$ in the analysis.

Assumption 1 requires $\lambda_{\text{tgt}} \geq 4\|A^*z\|^*$. This assumption on the regularization parameter is a standard assumption that is used in the literature to provide optimal bounds for recovery error [CP11, CT07, NRWY12]. The lower bound on λ_{tgt}, ensures $\gamma \leq \frac{5+4\delta}{3-4\delta}$. If we choose δ and η, we can set $\delta' = (1+\delta)\eta - 1$ to ensure that it satisfies (2.31). The parameter δ is directly related to satisfiability of (2.25) in Assumption 1. For example, if $\delta = 1/12$, then

$\gamma \leq 2$ and Assumption 1 is satisfied with $r = 2c$ if:

$$\frac{\rho_- \left(A, 9ck_0\right)}{\rho_+ \left(A, 432c^2k_0\gamma_{\text{inc}}\right)} > \frac{1}{2},$$
$$\rho_- \left(A, 432c^2k_0\gamma_{\text{inc}}\right) > 0.$$

Theoretically, the optimal choice of δ maximizes κ subject to existence of $r > 1$ that satisfies (2.25) and (2.26). In appendix A, we provide an upper bound on the number of measurement needed for (2.25) and (2.26) to be satisfied with high probability for given δ and $r > 1$ whenever rows of A are sub-Gaussian random vectors. The parameter η should be chosen to be greater than $\frac{1}{2}$ for (2.31) to be satisfied.

2.4.2 Convergence proof

The main part of the proof of Theorems 3 and 4 is establishing the fact that $K\left(x^{(t)}\right) \leq \tilde{k}$. Given that $K\left(x^{(t)}\right) \leq \tilde{k}$ for all t, Proposition 1 ensures that hypotheses of Proposition 2, i.e., strong convexity and gradient Lipschitz continuity over a restricted set, are satisfied. We adapt the same strategy as in [XZ13] and prove that $K\left(x^{(t)}\right) \leq \tilde{k}$ in a series of three lemmas. We have written the statement of the lemmas here, while their proofs are given in Appendix B. Lemma 1 states that if $\omega_\lambda(x)$ does not exceed a small fraction of λ, then x is close to x_0.

Lemma 1. *If* $\omega_\lambda(x) \leq \delta\lambda$ *and* $\rho_- \left(A, c\left(1+\gamma\right)^2 k_0\right) > 0$, *then:*

$$\max\left\{\|x - x_0\|, \frac{1}{\delta\lambda}\left(\phi_\lambda\left(x\right) - \phi_\lambda\left(x_0\right)\right)\right\} \leq \frac{ck_0\left(1+\gamma\right)\left(\left(1+\delta\right)\lambda + \|A^*z\|^*\right)}{\rho_-\left(A, c\left(1+\gamma\right)^2 k_0\right)}. \tag{2.34}$$

Note that if $\lambda \geq 4\|A^*z\|^*$ and $\delta \leq \frac{1}{4}$, we can simplify the conclusion of Lemma 1 as

$$\max\left\{\|x - x_0\|, \frac{1}{\delta\lambda}\left(\phi_\lambda\left(x\right) - \phi_\lambda\left(x_0\right)\right)\right\} \leq \frac{3ck_0\lambda\left(1+\gamma\right)}{2\rho_-\left(A, c\left(1+\gamma\right)^2 k_0\right)}$$

While the hypotheses of this lemma are true in the first step of every outer iteration of homotopy algorithm, $\omega_\lambda(x^{(t)})$ may not be decreasing in proximal-gradient algorithm. However, the objective decreases after every iteration of the proximal-gradient algorithm. Thus

to conclude that $x^{(t)}$ is close to x_0 in all the inner proximal-gradient steps we can use the following lemma:

Lemma 2. *Suppose Assumption 1 holds true, and $\lambda \geq \lambda_{\text{tgt}}$. If*

$$\phi_\lambda(x) - \phi_\lambda(x_0) \leq \frac{3ck_0\delta\lambda^2(1+\gamma)}{2\rho_-\left(A, c(1+\gamma)^2 k_0\right)},$$

then

$$\max\{\frac{1}{2\lambda}\|A(x-x_0)\|_2^2, \|x-x_0\|\} \leq \frac{9ck_0\lambda(1+\gamma)}{2\rho_-\left(A, c(1+\gamma)^2 k_0\right)}.$$

The proofs of Lemma 1 and Lemma 2 generalize the proofs of the corresponding lemmas in [XZ13] given for ℓ_1 norm to norms that satisfy Condition 2 using the structure of $\partial\|x_0\|$ given by (2.9). The last lemma provides an upper bound on $K(x^+)$, where x^+ is produced via a proximal-gradient step on x, as long as x satisfies the conclusion of Lemma 2 and Assumption 1 holds. The proof of Lemma 3 uses a slightly different approach than the one given in [XZ13] resulting in a simpler requirement on $\tilde{k}$ in Assumption 1.

Lemma 3. *Let $x^+ = \text{Prox}_{\lambda,L}(x)$ and suppose Assumption 1 holds true, and $\lambda \geq \lambda_{\text{tgt}}$. If $L \leq \gamma_{\text{inc}}\rho_+\left(A, 2\tilde{k}\right)$ and*

$$\max\{\frac{1}{2\lambda}\|A(x-x_0)\|_2^2, \|x-x_0\|\} \leq \frac{9ck_0\lambda(1+\gamma)}{2\rho_-\left(A, c(1+\gamma)^2 k_0\right)},$$

then $K(x^+) \leq \tilde{k}$.

2.4.3 Proof of Theorem 3

First we show that $L_t \leq \gamma_{\text{inc}}\rho_+\left(A, 2\tilde{k}\right)$ and $K\left(x^{(t)}\right) \leq \tilde{k}$ for all $t \geq 0$. The inequalities hold true for $t = 0$ by the hypothesis. Suppose $L_t \leq \gamma_{\text{inc}}\rho_+\left(A, \tilde{k}\right)$ and $K\left(x^{(t)}\right) \leq \tilde{k}$ for some $t \geq 0$. Since $\phi_\lambda\left(x^{(t)}\right) \leq \phi_\lambda\left(x^{(0)}\right)$, by Lemma 2, we have:

$$\max\{\frac{1}{2\lambda}\left\|A\left(x^{(t)}-x_0\right)\right\|_2^2, \left\|x^{(t)}-x_0\right\|\} \leq \frac{9ck_0\lambda(1+\gamma)}{2\rho_-\left(A, c(1+\gamma)^2 k_0\right)}.$$

By Lemma 2, Lemma 3 and Theorem 2, for any $L \leq \gamma_{\text{inc}}\rho_+\left(A, 2\tilde{k}\right)$

$$K\left(\text{Prox}_{\lambda,L}\left(x^{(t)}\right)\right) \leq \tilde{k},$$
$$K\left(\text{Prox}_{\lambda,L}\left(x^{(t)}\right) - x^{(t)}\right) \leq 2\tilde{k}.$$

Now we can use Proposition 1 to conclude that $M_{t+1} \leq \gamma_{\text{inc}}\rho_+\left(A, 2\tilde{k}\right)$ hence $L_{t+1} \leq M_{t+1}/\gamma_{\text{dec}} \leq \gamma_{\text{inc}}\rho_+\left(A, 2\tilde{k}\right)$. In addition, by Lemma 3, $K\left(x^{(t+1)}\right) = K\left(\text{Prox}_{\lambda,M_{t+1}}\left(x^{(t)}\right)\right) \leq \tilde{k}$.

Since $\text{Prox}_{\lambda,L}(x^*) = x^*$ for any $L > 0$, by Lemmas 1, 2, and 3, $K\left(x^*\right) \leq \tilde{k}$. By Theorem 2, we have:

$$K\left(x^{(t+1)} - x^{(t)}\right) \leq 2\tilde{k},\ K\left(x^{(t)} - x^*\right) \leq 2\tilde{k},$$

which yields

$$\begin{aligned}\left\|A^*A\left(x^{(t+1)} - x^{(t)}\right)\right\|^* &= \max_{a\in\mathcal{G}_{\|\cdot\|}} \langle a, A^*A\left(x^{(t+1)} - x^{(t)}\right)\rangle \\ &= \max_{a\in\mathcal{G}_{\|\cdot\|}} \langle Aa, A\left(x^{(t+1)} - x^{(t)}\right)\rangle \leq \sqrt{\rho_+\left(A,1\right)\rho_+\left(A, 2\tilde{k}\right)}\left\|x^{(t+1)} - x^{(t)}\right\|_2.\end{aligned} \tag{2.35}$$

Now Proposition 1 and (2.35) ensure that all the hypotheses of Proposition 2 are satisfied with $\mu_f = \rho_-\left(A, 2\tilde{k}\right)$, $L_f = \rho_+\left(A, 2\tilde{k}\right)$. $L'_f = \sqrt{\rho_+\left(A,1\right)\rho_+\left(A, 2\tilde{k}\right)}$ and $\theta = 1$. Thus the conclusion follows from Proposition 2.

2.4.4 Proof of Theorem 4

Let $y_t^* \in \text{argmin}\,\phi_{\lambda_t}(y)$. For the ease of notation let $\lambda_{N+1} \leftarrow \lambda_{\text{tgt}}$. First we show that $\omega_{\lambda_{t+1}}\left(y^{(t)}\right) \leq \delta\lambda_{t+1}$ and $K\left(y^{(t)}\right) \leq \tilde{k}$ for $t = 0, 1, \ldots, N$. When $t = 0$, we have $y^{(0)} = 0$ and

$\lambda_0 = \|A^*b\|^*$. Therefore, $K(y^{(0)}) = 0$ and

$$\begin{aligned}\omega_{\lambda_1}(y^{(0)}) &= \min_{\xi \in \partial\|0\|} \|A^*b + \lambda_1 \xi\|^* \\ &\text{Since } \frac{-A^*b}{\lambda_0} \in \partial\|0\| \\ &\leq \left\| A^*b - \frac{\lambda_1}{\lambda_0} A^*b \right\|^* \\ &= (1-\eta)\lambda_0 \leq \delta\lambda_1,\end{aligned}$$

where in the last inequality we used (2.31). Suppose $\omega_{\lambda_t}\left(y^{(t-1)}\right) \leq \delta\lambda_t$ and $K\left(y^{(t-1)}\right) \leq \tilde{k}$. By Theorem 3, we have:

$$K\left(y^{(t)}\right) \leq \tilde{k}.$$

By (2.17), the stopping condition in the proximal gradient algorithm ensures $\omega_{\lambda_t}\left(y^{(t)}\right) \leq \delta'\lambda_t$. Therefore, there exists $\xi \in \partial\left\|y^{(t)}\right\|$ such that $\left\|A^*\left(Ay^{(t)} - b\right) + \lambda_t\xi\right\|^* \leq \delta'\lambda_t$. Now using hypothesis (2.31), we get:

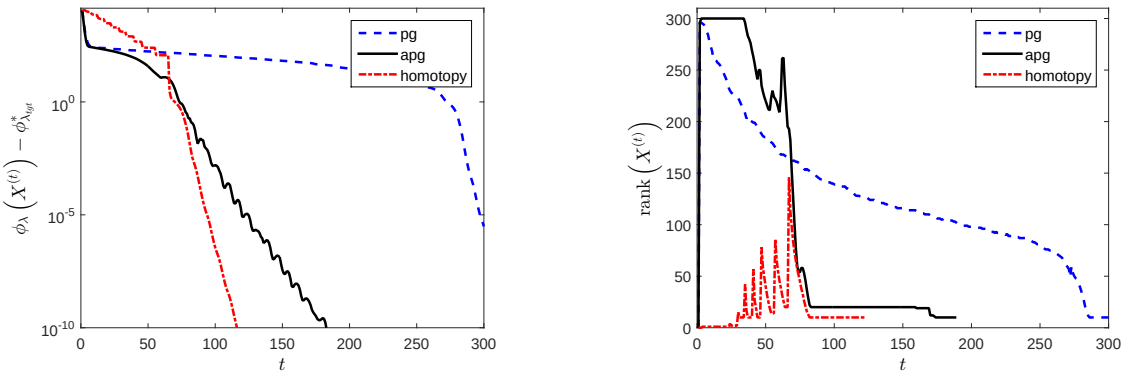
$$\begin{aligned}\omega_{\lambda_{t+1}}\left(y^{(t)}\right) &\leq \left\|A^*\left(Ay^{(t)} - b\right) + \lambda_{t+1}\xi\right\|^* \\ &\leq \left\|A^*\left(Ay^{(t)} - b\right) + \lambda_t\xi\right\|^* + \|(\lambda_{t+1} - \lambda_t)\,\xi\|^* \\ &\leq \omega_{\lambda_t}\left(y^{(t)}\right) + (\lambda_t - \lambda_{t+1}) \leq \left(-1 + (\delta'+1)/\eta\right)\lambda_{t+1} \leq \delta\lambda_{t+1}.\end{aligned}$$

By Lemma 1 and the comment that follows it, for all $t = 0, \ldots, N$, we have

$$\begin{aligned}\left\|y^{(t)} - y^*_{t+1}\right\| &\leq \left\|y^{(t)} - x_0\right\| + \left\|y^*_{t+1} - x_0\right\| \\ &\leq \frac{ck_0\,(1+\gamma)\left((2+\delta)\,\lambda_{t+1} + 2\|A^*z\|^*\right)}{\rho_-\left(A, c\,(1+\gamma)^2\,k_0\right)} \\ &\leq \frac{3ck_0\,(1+\gamma)\,\lambda_{t+1}}{\rho_-\left(A, c\,(1+\gamma)^2\,k_0\right)}.\end{aligned}$$

Hence

$$\begin{aligned}\phi_{\lambda_{t+1}}\left(y^{(t)}\right) - \phi_{\lambda_{t+1}}\left(y^*_{t+1}\right) &\leq \langle \omega_{\lambda_{t+1}}\left(y^{(t)}\right), y^{(t)} - y^*_{t+1}\rangle \\ &\leq \omega_{\lambda_{t+1}}\left(y^{(t)}\right)\left\|y^{(t)} - y^*_{t+1}\right\| \\ &\leq \frac{3\delta ck_0\,(1+\gamma)\,\lambda^2_{t+1}}{\rho_-\left(A, c\,(1+\gamma)^2\,k_0\right)}.\end{aligned}$$



(A) Objective gap vs. iteration

(B) Rank vs. iteration

Figure 2.1: Comparison of homotopy, proximal-gradient and accelerated proximal-gradient algorithms for problem 1

Now the upper bounds in (2.32) and (2.33) on the number of inner iterations follow from the second conclusion in Theorem 3.

By (B.21), we have

$$\|y-y^*\| \leq \|y-y_0\| + \|y_0-y^*\| \leq \frac{9ck_0\lambda_{\mathrm{tgt}}\left(1+\gamma\right)}{\rho_-\left(A, c\left(1+\gamma\right)^2 k_0\right)}.$$

By convexity of $\phi_{\lambda_{\mathrm{tgt}}}$, we get:

$$\begin{aligned}\phi_{\lambda_{\mathrm{tgt}}}\left(y\right)-\phi_{\lambda_{\mathrm{tgt}}}\left(y^*\right) &\leq \left\langle \omega_{\lambda_{tgt}}\left(y\right), y-y^*\right\rangle \\ &\leq \frac{9ck_0\lambda_{\mathrm{tgt}}\left(1+\gamma\right)\epsilon}{\rho_-\left(A, c\left(1+\gamma\right)^2 k_0\right)}.\end{aligned}$$

2.5 Numerical Experiments

We consider two problems. The details of each problem are summarized in the following table:

	Problem 1	Problem 2
Objective	$\frac{1}{2}\|A\,\mathrm{vec}(X)+b\|_2^2+\lambda\|X\|_*$	$\frac{1}{2}\|A\,\mathrm{vec}(X)+b\|_2^2+\lambda\|X\|_{1,2}$
dimension of X_0	300×300	50×1000
$K(X_0)$	$\mathrm{rank}(X_0)=10$	# of non-zero columns of $X_0=50$
#of samples	$m=20000$	$m=18000$
b	$A\,\mathrm{vec}(X_0)+z$	$A\,\mathrm{vec}(X_0)+z$
$A_{i,j}$ sampled from	$\mathcal{N}(0,1/\sqrt{m})$	$\{-1/\sqrt{m},1/\sqrt{m}\}$ uniformly at rand.
z_i sampled from	$\mathcal{U}(-0.005,0.005)$	$\mathcal{U}(-0.005,0.005)$

In the homotopy algorithm, $\lambda_0=\left\|A^Tb\right\|^*$ and $\lambda_{\mathrm{tgt}}=4\left\|A^Tz\right\|^*$, while in the proximal-gradient algorithm $\lambda=\lambda_{\mathrm{tgt}}$. The default values of η and δ' in the homotopy algorithm are $\eta=0.6$, $\delta'=0.2$.

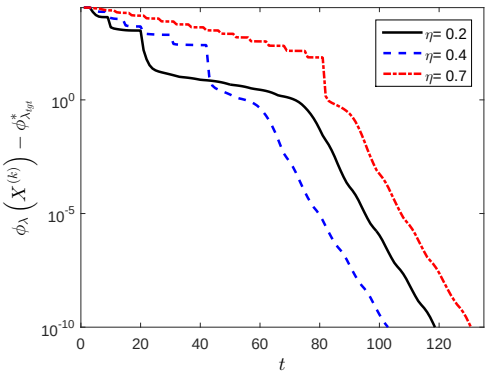
Problem 1. Figure 2.1 demonstrates the overall linear rate of convergence of proximal-gradient homotopy algorithm (homotopy) applied to this problem and compares it with proximal-gradient algorithm (PG) and its accelerated version (APG). As rank vs. iteration plot demonstrates, the proximal-gradient algorithm speeds up to a linear rate when the rank drops to a certain level, while the homotopy algorithm keeps the rank at a level that ensures a linear rate of convergence.

We examine the performance of homotopy algorithm with three different values of η and δ' in Figure 2.2. For η to satisfy the condition of Theorem 4, it is necessary that $\eta>0.5$. However, as Figure 2.2 demonstrates, one can choose $\eta\leq 0.5$ and still get an overall linear rate of convergence. For example, when $\eta=0.2$, at the beginning of the last stage where $\lambda=\lambda_{\mathrm{tgt}}$, $X^{(k)}$ is not low-rank and the algorithm has a sublinear rate of convergence, but nevertheless the algorithm converges faster with $\eta=0.2$ than $\eta=0.7$. Homotopy algorithm appears to be even less sensitive to δ'. As δ' gets closer to 1, the rank of $X^{(k)}$ jumps higher, which can cause a slowdown in convergence specially at the beginning of each stage.

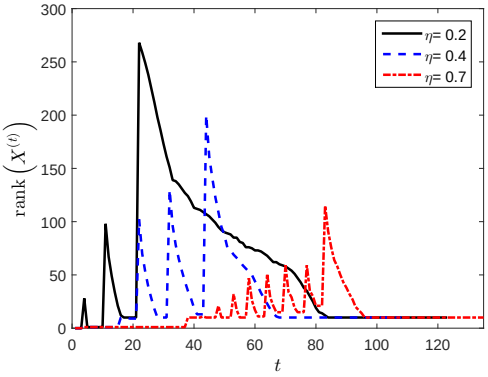
In Figure 2.3A, we have compared recovery error of the following algorithms: SVP, FPC, APGL, homotopy, proximal-gradient and its accelerated version. In SVP we provide the

algorithm with the rank of X_0, while in SVP2 we use the same heuristic that is proposed in [JMD10] to estimate the rank (other algorithms do not receive the rank of X_0).We have implemented the FPC algorithm with the backtracking procedure which improves the performance of the algorithm. Both APGL and APGL2 have been implemented with continuation over λ with the latter utilizing an extra truncation heuristic proposed in [TY10]. The method of continuation for APGL is the same as the one proposed in [TY10]; we reduce λ by a factor of 0.7 after three iterations or whenever the stopping criterion is met whichever comes first. In FPC and APGL similar to the homotopy algorithms, $\lambda_0 = \left\|A^T(b)\right\|^*$ and $\lambda_{\text{tgt}} = 4\left\|A^T(z)\right\|^*$. We have used the default values of the parameters in all the algorithms. Note that APGL2 has an extra truncation procedure which improves the recovery error. Finally, Figure 2.3B shows the objective gap for the algorithms for which the quantity is meaningful.

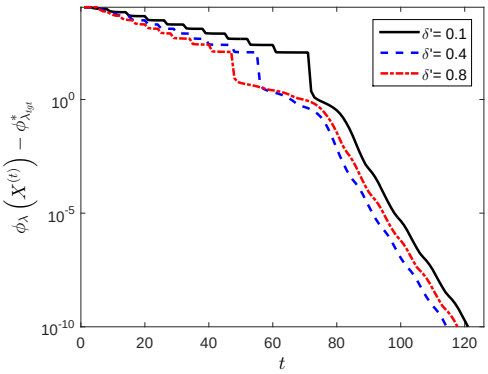
Problem 2. Figure 2.4 demonstrates the linear convergence of homotopy algorithm for this problem and compares the performance with that of proximal-gradient algorithm and its accelerated version. Similar to problem 1, homotopy algorithm keeps the number of non-zero columns below a certain level. In homotopy algorithm $\delta' = 0.2$ and $\eta = 0.6$.



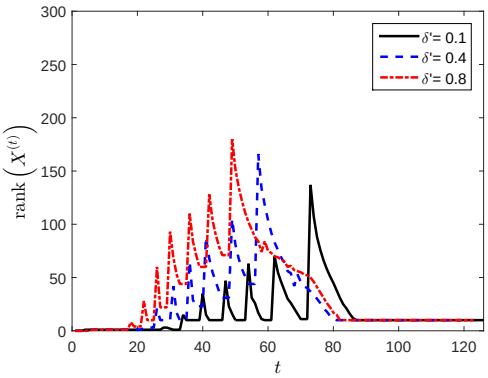
(A) Objective gap vs. iteration



(B) Rank vs. iteration



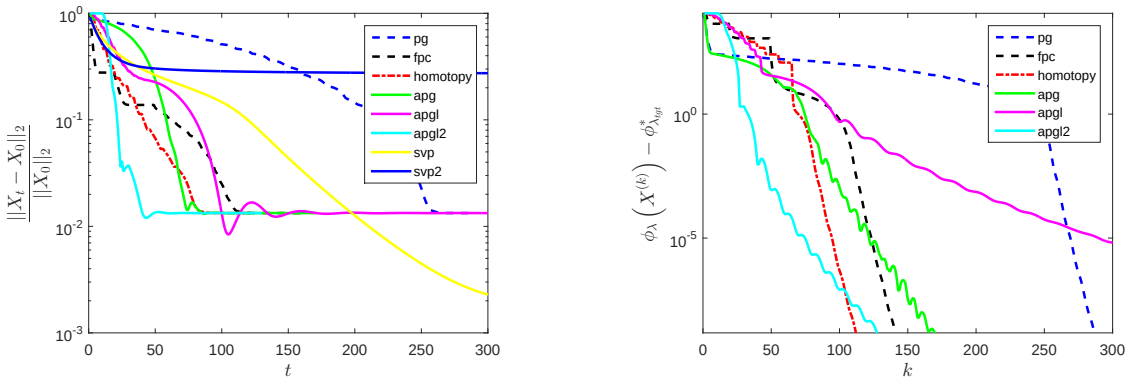
(C) Objective gap vs. iteration



(D) Rank vs. iteration

Figure 2.2: (a), (b): Performance of homotopy algorithm with $\delta' = 0.2$ and three different values of η,

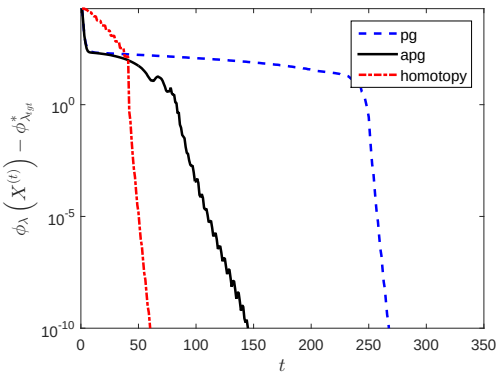
(c), (d): Performance of homotopy algorithm with $\eta = 0.6$ and three different values of δ'



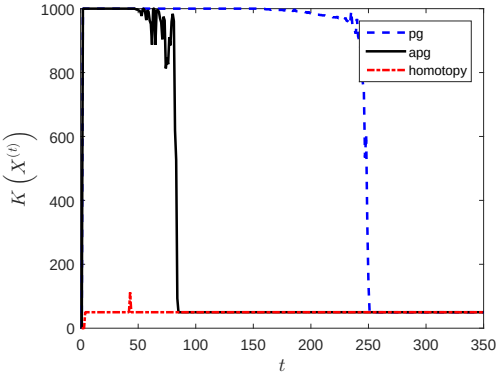
(A) Recovery error vs. iteration

(B) Objective gap vs. iteration

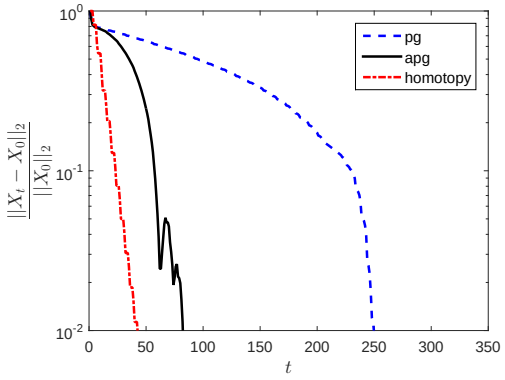
Figure 2.3: Comparison between SVP, FPC, APGL, homotopy, proximal-gradient and its accelerated version



(A) Objective gap vs. iteration



(B) Number of non-zero columns vs. iteration



(C) Recovery error vs. iteration

Figure 2.4: Comparison of homotopy, proximal-gradient and accelerated proximal-gradient algorithms for problem 2

Chapter 3

ONLINE CONIC OPTIMIZATION

Given a proper convex cone $K \subset \mathbb{R}^n$, let $\psi : K \mapsto \mathbb{R}$ be an upper semi-continuous concave function. Consider the optimization problem

$$\begin{aligned} \text{maximize} \quad & \psi\left(\sum_{t=1}^{m} A_t x_t\right) \\ \text{subject to} \quad & x_t \in F_t, \quad \forall t \in [m], \end{aligned} \tag{3.1}$$

where for all $t \in [m] := \{1, 2, \ldots, m\}$, $x_t \in \mathbb{R}^k$ are the optimization variables and F_t are compact convex constraint sets. We assume $A_t \in \mathbb{R}^{n \times k}$ maps F_t to K; for example, when $K = \mathbb{R}^n_+$ and $F_t \subset \mathbb{R}^k_+$, this assumption is satisfied if A_t has nonnegative entries. We consider problem (3.1) in the online setting, where it can be viewed as a sequential game between a player (online algorithm) and an adversary. At each step t, the adversary reveals A_t, F_t and the algorithm chooses $\hat{x}_t \in F_t$. The performance of the algorithm is measured by its competitive ratio, i.e., the ratio of objective value at $\hat{x}_1, \ldots, \hat{x}_m$ to the offline optimum.

Problem (3.1) covers (convex relaxations of) various online combinatorial problems including online bipartite matching [KVV90], the "adwords" problem [MSVV07], and the secretary problem [Kle05]. More generally, it covers online linear programming (LP) [BN09], online packing/covering with convex cost [ACP14, BCG$^+$14, CHK15], and generalization of adwords [DJ12]. Online LP is an important example on the positive orthant:

$$\begin{aligned} \text{maximize} \quad & \sum_{t=1}^{m} c_t^T x_t \\ \text{subject to} \quad & \sum_{t=1}^{m} B_t x_t \leq b \\ & \mathbf{1}^T x_t \leq 1,\ x_t \in \mathbf{R}^k_+, \quad \forall t \in [m]. \end{aligned}$$

Here $F_t = \{x \in \mathbf{R}^k_+ \mid \mathbf{1}^T x \leq 1\}$, $A_t^T = [c_t, B_t^T]$ and $\psi(u, v) = u + I_{\{\cdot \leq b\}}(v)$ where $I_{\{\cdot \leq b\}}(v)$ is the concave indicator function of the set $\{v \leq b\}$.

For a class of examples on the PSD cone S^n_+ consider the following online game: At round t, the adversary picks a matrix $B_t \in S^n_+$ and a scalar $c_t > 0$. The online algorithm chooses $x_t \in [0, c_t]$. The goal of the algorithm is to maximize $H\left(\sum_{t=1}^m B_t x_t\right)$ subject to the budget constraint $\sum_{t=1}^m x_t \leq b$. The offline problem can be written as:

$$\text{maximize} \quad H\left(\sum_{t=1}^m B_t x_t\right) \qquad \text{subject to} \quad \begin{cases} \sum_{t=1}^m x_t \leq b, \\ 0 \leq x_t \leq c_t, \;\; t = 1, \ldots, m. \end{cases} \tag{3.2}$$

The main examples we study are *online experiment design* (a classic problem in statistics) and *online graph formation*. In both problems, $B_t = a_t a_t^T$ for some $a_t \in \mathbf{R}^n_+$. In online experiment design or sensor selection, the vector a_t is an experiment or measurement vector that provides linear noisy measurements from a vector w, i.e., $\xi_t = \langle a_t, w \rangle + n_t$, where $n_t \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise. The algorithm makes online decisions on whether to pick or drop a measurement vector. The goal is to minimize error covariance of the *maximum a priori estimator (MAP)* of w. The convex relaxation of this problem (where the binary decision variables is relaxed to a continuous one) can be expressed as problem (3.1) with H given by

$$\begin{aligned} H(U) &= \log\det(\epsilon I + U) \quad \text{D-optimal criterion,} \\ H(U) &= -\mathbf{tr}(\epsilon I + U)^{-1} \quad \text{A-optimal criterion,} \\ H(U) &= -\mathbf{tr}(\epsilon I + U)^{-p} \quad \text{p}^{th} \text{ mean criterion,} \end{aligned}$$

whence the prior distribution on w is $\mathcal{N}\left(0, \frac{1}{\epsilon} I\right)$.

The competitive performance of online algorithms has been studied mainly under the worst-case model (e.g., in [MSVV07]) or stochastic models (e.g., in [Kle05]). In the worst-case model one is interested in lower bounds on the competitive ratio that hold for any $(A_1, F_1), \ldots, (A_m, F_m)$. In stochastic models, adversary chooses a probability distribution from a family of distributions to generate $(A_1, F_1), \ldots, (A_m, F_m)$, and the competitive ratio is

calculated using the expected value of the algorithm's objective value. The two most studied stochastic models are random permutation and i.i.d. models. In the random permutation model the adversary is limited to choose distributions that are invariant under the random permutation while in the i.i.d. model $(A_1, F_1), \ldots, (A_m, F_m)$ are required to be independent and identically distributed. Note that the worst case model can also be viewed as a stochastic model if one allows distributions with singleton support.

Online bipartite matching, and its generalization the "adwords" problem, are two main examples that have been studied under the worst case model. The greedy algorithm achieves a competitive ratio of $1/2$ while the optimal algorithm achieves a competitive ratio of $1 - 1/e$ as bid to budget ratio goes to zero [MSVV07, BJN07, KVV90, KP00]. A more general version of adwords in which each agent (advertiser) has a concave cost has been studied in [DJ12]. On the other hand, secretary problem and its generalizations are stated for random permutation model since the worst case competitive ratio of any online algorithm can be arbitrary small [BIKK08, Kle05]. The convex relaxation of the online secretary problem is a simple linear program (LP) with one constraint. Therefore, the competitive ratio of online algorithms for LP has either been analyzed under random permutation or i.i.d models [AWY09, FHK$^+$10, DJSW11, JL12, KRTV14] or under the worst case model with further restrictions on the problem data [BN09]. Several authors have also proposed algorithms for adwords and bipartite matching problems under stochastic models which have a better competitive ratio than the competitive ratio under the worst case model [MY11, KMT11, FMMM09, MGS12, JL13, DH09].

The majority of algorithms proposed for the problems mentioned above rely on a primal-dual framework [BJN07, BN09, ACP14, DJ12, BCG$^+$14]. The differentiating point among the algorithms is the method of updating the dual variable at each step, since once the dual variable is updated the primal variable can be assigned using a simple assignment rule based on complementary slackness condition. A simple and efficient method of updating the dual variable is through a first order online learning step. For example, the algorithm stated in [DJSW11] for online linear programming uses mirror descent with entropy regularization

(multiplicative weight updates algorithm) once written in the primal dual language. Recently, the work in [DJSW11] was independently extended to the random permutation model in [GM14, ESF14, AD14]. In [AD14], the authors provide competitive difference bound for online convex optimization under random permutation model as a function of the regret bound for the online learning algorithm applied to the dual.

In this chapter, we consider two versions of the greedy algorithm for problem (3.1) The first algorithm, Algorithm 6, updates the primal and dual variables sequentially. The second algorithm, Algorithm 3, provides a direct saddle-point representation of what has been described informally in the literature as "continuous updates" of primal and dual variables. This saddle point representation allows us to generalize this type of updates to non-smooth functions. In section 3.2, we bound the competitive ratios of the two algorithms. A sufficient condition (*diminishing returns (DR)*) on the objective function that guarantees a non-trivial worst case competitive ratio is introduced. We show that the competitive ratio is at least $\frac{1}{2}$ for a monotone non-decreasing objective function. Examples that satisfy the sufficient condition (on the positive orthant and the positive semidefinite cone) are given. In section 3.3, we derive optimal algorithms, as variants of the greedy algorithm applied to a smoothed version of ψ. We require the smooth version of ψ to satisfy the DR assumption. Nesterov smoothing provides an optimal algorithm for online LP. The main contribution of this chapter is to show how one can derive the optimal smoothing function (or from the dual point of view the optimal regularization function) for separable ψ on positive orthant and for trace functions on the PSD cone by solving a convex optimization problem. In the case of separable functions on the positive orthant this gives a implementable algorithm that achieves the optimal competitive ratio derived in [DJ12]. We also show how this convex optimization can be modified for the design of the smoothing function specifically for the sequential algorithm. In contrast, [DJ12] only considers continuous updates algorithm, which we show can be derived form Algorithm 3.

In the case of trace functions over the PSD cone, to impose the DR property on the smooth function, we use the Löwner's theorem, characterizing operator monotone functions, to im-

pose this DR property in a computationally effective way by requiring the smooth function to have a certain integral representation (see (3.20)). This allows us to provide competitive ratio bounds for problems such as A-optimal experiment design where the objective function does not satisfy the DR property.

Notation. We denote the transpose of a matrix A by A^T. The inner product on $\mathbb{R}^n$ is denoted by $\langle \cdot, \cdot \rangle$. The eigenvalues of a symmetric $n \times n$ matrix U are denoted by $\lambda_1(U), \ldots, \lambda_n(U)$. We use $(\cdot)_+$ to denote the positive part, defined as $(u)_+ = \max\{0, u\}$.

Given a function $\psi : \mathbb{R}^n \mapsto \mathbb{R}$, ψ^* denotes the concave conjugate of ψ defined as

$$\psi^*(y) = \inf_u \langle y, u \rangle - \psi(u),$$

for all $y \in \mathbb{R}^n$. For a concave function ψ, $\partial\psi(u)$ denotes the set of supergradients of ψ at u, i.e., the set of all $y \in \mathbb{R}^n$ such that

$$\forall u' \in \mathbb{R}^n : \quad \psi(u') \leq \langle y, u' - u \rangle + \psi(u).$$

The set $\partial\psi$ is related to the concave conjugate function ψ^* as follows. For an upper semi-continuous concave function ψ we have

$$\partial\psi(u) = \operatorname*{argmax}_y \langle y, u \rangle - \psi^*(y).$$

Under this condition, $(\psi^*)^* = \psi$.

A differentiable function ψ has a Lipschitz continuous gradient with respect to $\|\cdot\|$ with continuity parameter $\frac{1}{\mu} > 0$ if for all $u, u' \in \mathbb{R}^n$,

$$\|\nabla\psi(u') - \nabla\psi(u)\|^* \leq \frac{1}{\mu}\|u - u'\|,$$

where $\|\cdot\|^*$ is the dual norm to $\|\cdot\|$. For an upper semi-continuous concave function ψ, this is equivalent to ψ^* being μ-strongly concave with respect to $\|\cdot\|^*$ (see, for example, [RWW98, chapter 12]).

The dual cone K^* of a cone $K \subset \mathbb{R}^n$ is defined as $K^* = \{y \mid \langle y, u \rangle \geq 0 \ \ \forall u \in K\}$. Two examples of self-dual cones are the positive orthant $\mathbb{R}^n_+$ and the cone of $n \times n$ positive

semidefinite matrices S_+^n. A proper cone (pointed convex cone with nonempty interior) K induces a partial ordering on $\mathbb{R}^n$ which is denoted by $\leq_K$ and is defined as

$$x \leq_K y \Leftrightarrow y - x \in K.$$

For two sets $F, G \subset \mathbb{R}^n$, we write $F \leq_K G$ when $u \leq_K v$ for all $u \in F, v \in G$.

3.1 Two greedy algorithms

The (Fenchel) dual problem for problem (3.1) is given by

$$\text{minimize} \quad \sum_{t=1}^{m} \sigma_t(A_t^T y) - \psi^*(y), \tag{3.3}$$

where the optimization variable is $y \in \mathbb{R}^n$, and σ_t denotes the *support function* for the set F_t defined as $\sigma_t(z) = \sup_{x \in F_t} \langle x, z \rangle$. We denote the optimal dual objective with $D^\star$.

A pair $(x^*, y^*) \in (F_1 \times \dots F_m) \times K^*$ is an optimal primal-dual pair if and only if

$$x_t^* \in \operatorname*{argmax}_{x \in F_t} \langle x, A_t^T y^* \rangle \quad \forall t \in [m], \tag{3.4}$$

$$y^* \in \partial\psi(\sum_{t=1}^{m} A_t x_t^*) \tag{3.5}$$

Based on these optimality conditions, we consider two algorithms. Algorithm 6 updates the primal and dual variables *sequentially*, by maintaining a dual variable $\hat{y}_t$ and using it to assign $\hat{x}_t \in \operatorname{argmax}_{x \in F_t} \langle x, A_t^T \hat{y}_t \rangle$. The algorithm then updates the dual variable based on the second optimality condition (3.5) [1]. By the assignment rule, we have $A_t \hat{x}_t \in \partial\sigma_t(\hat{y}_t)$, and the dual variable update can be viewed as

$$\hat{y}_{t+1} \in \operatorname*{argmin}_y \langle \sum_{s=1}^{t} A_s \hat{x}_s, y \rangle - \psi^*(y).$$

Therefore, the dual update is the same as the update in dual averaging [Nes09] or Follow The Regularized Leader (FTRL) [SSS07b, AHR08][SS11, section 2.3] algorithm with regularization $-\psi^*(y)$.

[1] we assume that $\partial\psi$ exists on K, otherwise, we can modify the algorithm to start from a point in the interior of the cone very close to zero.

Algorithm 2 Sequential Update

Initialize $\hat{y}_1 \in \partial\psi(0)$

for $t \leftarrow 1$ to m **do**

Receive A_t, F_t

$\hat{x}_t \in \text{argmax}_{x \in F_t} \langle x, A_t^T \hat{y}_t \rangle$

$\hat{y}_{t+1} \in \partial\psi(\sum_{s=1}^{t} A_s \hat{x}_s)$

end for

Algorithm 3 updates the primal and dual variables *simultaneously*, ensuring that

$$\tilde{x}_t \in \operatorname*{argmax}_{x \in F_t} \langle x, A_t^T \tilde{y}_t \rangle, \qquad \tilde{y}_t \in \partial\psi(\sum_{s=1}^{t} A_s \tilde{x}_s).$$

This algorithm is inherently more complicated than algorithm 6, since finding $\tilde{x}_t$ involves solving a saddle-point problem. This can be solved by a first order method like mirror descent algorithm for saddle point problems. In contrast, the primal and dual updates in algorithm 6 solve two separate maximization and minimization problems [2]

Algorithm 3 Simultaneous Update

for $t \leftarrow 1$ to m **do**

Receive A_t, F_t

$(\tilde{y}_t, \tilde{x}_t) \in \arg\min_y \max_{x \in F_t} \langle y, A_t x + \sum_{s=1}^{t-1} A_s \tilde{x}_s \rangle - \psi^*(y)$

end for

For a reader more accustomed to optimization algorithms, we would like to point to alternative views on these algorithms. If $0 \in F_t$ for all t, then an alternative view on Algorithm 3 is coordinate minimization. Initially, all the coordinates are set to zero, then at

[2] Also if the original problem is a convex relaxation of an integer program, meaning that each $F_t = \text{conv}\mathcal{F}_t$ where $\mathcal{F}_t \subset \mathbb{Z}^l$, then $\hat{x}_t$ can always be chosen to be integral while integrality may not hold for the solution of the second algorithm.

step t

$$\tilde{x}_t \in \operatorname*{argmax}_{x \in F_t} \psi \left(\sum_{s=1}^{t} A_t \tilde{x}_s + A_t x \right). \tag{3.6}$$

If in addition ψ is differentiable, Algorithm 6, at each time step t, applies one iteration of Frank-Wolfe algorithm [FW56] for solving (3.6).

3.2 Competitive ratio bounds and examples for ψ

In this section, we derive bounds on the competitive ratios of Algorithms 6 and 3 by bounding their respective duality gaps. Let $\tilde{y}_{m+1}$ to be the minimum element in $\partial\psi(\sum_{t=1}^{m} A_t \tilde{x}_t)$ with respect to ordering $\leq_{K^*}$ (such an element exists in the superdifferential by Assumption (2), which appears later in this section). We also choose $\hat{y}_{m+1}$ to be the minimum element in $\partial\psi(\sum_{t=1}^{m} A_t \hat{x}_t)$ with respect to $\leq_{K^*}$. Note that $\hat{y}_{m+1}$ is used for analysis and does not play a role in the assignments of Algorithm 6. Let P_{seq} and P_{sim} denote the primal objective values for the primal solution produced by the algorithms 6 and 3, and D_{seq} and D_{sim} denote the corresponding dual objective values,

$$P_{\text{seq}} = \psi \left(\sum_{t=1}^{m} A_t \hat{x}_t \right), \qquad P_{\text{sim}} = \psi \left(\sum_{t=1}^{m} A_t \tilde{x}_t \right),$$
$$D_{\text{seq}} = \sum_{t=1}^{m} \sigma_t(A_t^T \hat{y}_t) - \psi^*(\hat{y}_{m+1}), \qquad D_{\text{sim}} = \sum_{t=1}^{m} \sigma_t(A_t^T \tilde{y}_t) - \psi^*(\tilde{y}_{m+1}).$$

The next lemma provides a lower bound on the duality gaps of both algorithms.

Lemma 4. *The duality gap for Algorithm 3 is lower bounded as*

$$P_{\text{sim}} - D_{\text{sim}} \geq \psi^*(\tilde{y}_{m+1}) + \psi(0),$$

and the duality gap for Algorithm 6 is lower bounded as

$$P_{\text{seq}} - D_{\text{seq}} \geq \psi^*(\hat{y}_{m+1}) + \psi(0) + \sum_{t=1}^{m} \langle A_t \hat{x}_t, \hat{y}_{t+1} - \hat{y}_t \rangle.$$

Furthermore, if ψ has a Lipschitz continuous gradient with parameter $1/\mu$ with respect to $\|\cdot\|$, then

$$P_{\text{seq}} - D_{\text{seq}} \geq \psi^*(\hat{y}_{m+1}) + \psi(0) - \sum_{t=1}^{m} \frac{1}{2\mu}\|A_t\hat{x}_t\|^2. \tag{3.7}$$

Note that right hand side of (3.7) is exactly the regret bound of the FTRL algorithm (with a negative sign) [SSS07a]. The proof is given in Appendix C.1. To simplify the notation in the rest of this chapter, we assume $\psi(0) = 0$.

Now we state a sufficient condition on ψ that leads to non-trivial competitive ratios.One can interpret this assumption as having "diminishing returns" with respect to the ordering induced by a cone. Examples of functions that satisfy this assumption will appear later in this section.

Assumption 2. *Whenever $u \geq_K v$, there exists $y \in \partial\psi(u)$ that satisfies*

$$y \leq_{K^*} z,$$

for all $z \in \partial\psi(v)$.

When ψ is differentiable, Assumption 2 simplifies to

$$u \geq_K v \;\Rightarrow\; \nabla\psi(u) \leq_{K^*} \nabla\psi(v).$$

That is, the gradient, as a map from $\mathbb{R}^n$ (equipped with $\leq_K$) to $\mathbb{R}^n$ (equipped with $\leq_{K^*}$), is order-reversing (also known as *antitone*). If ψ satisfies Assumption 2, then so does $\psi \circ A$, when A is a linear map such that $\langle y, Av\rangle \geq 0$, for all $v \in K$ and $y \in K^*$. When ψ is twice differentiable, Assumption 2 is equivalent to $\langle w, \nabla^2\psi(u)v\rangle \leq 0$, for all $u, v, w \in K$. For example, this is equivalent to Hessian being element-wise non-positive when $K = \mathbf{R}^n_+$.

3.2.1 Competitive ratio for non-decreasing ψ

To quantify the competitive ratio of the algorithms, we define $\alpha_\psi : K \mapsto \mathbf{R}$ as

$$\alpha_\psi(u) = \sup\{c \mid \psi^*(y) \geq c\psi(u),\ \forall y \in \partial\psi(u)\}, \tag{3.8}$$

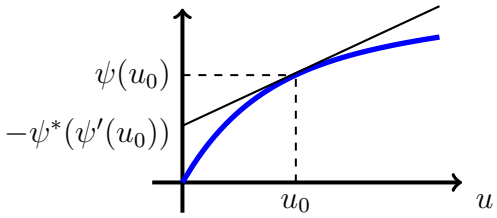


Figure 3.1: ψ' is the derivative of ψ and $-\psi^*(\psi'(u_0))$ is the y-intercept of the tangent to the function graph at u_0. $\alpha_\psi(u_0)$ is the ratio of $\psi^*(\psi'(u_0))$ to $\psi(u_0)$.

which for $u \neq 0$ can also be written as

$$\alpha_\psi(u) = \inf_{y \in \partial\psi(u)} \frac{\psi^*(y)}{\psi(u)}.$$

Since $\psi^*(y) + \psi(u) = \langle y, u \rangle$ for all $y \in \partial\psi(u)$, α_ψ is equivalent to[3]

$$\begin{aligned} \alpha_\psi(u) &= \sup\{c \mid \langle y, u \rangle \geq (c+1)\psi(u),\ \forall y \in \partial\psi(u)\} \\ &= \sup\{c \mid \psi'(u; u) \geq (c+1)\psi(u)\}, \end{aligned} \tag{3.9}$$

where $\psi'(u; v)$ is the directional derivative of ψ at u in the direction of v. Note that for any $u \in K$, we have $-1 \leq \alpha_\psi(u) \leq 0$ since for any $y \in \partial\psi(u)$, by concavity of ψ and the fact that $y \in K^*$, we have $0 \leq \langle y, u \rangle \leq \psi(u) - \psi(0)$. If ψ is a linear function then $\alpha_\psi = 0$, while if $0 \in \partial\psi(u)$ for some $u \in K$, then $\alpha_\psi(u) = -1$. Figure 3.1 shows a differentiable function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}$. The value of $\alpha(u_0)$ is the ratio of $\psi^*(\psi'(u_0))$ to $\psi(u_0)$ as shown.

The next theorem provides lower bounds on the competitive ratio of the two algorithms.

Theorem 5. *If Assumption 2 holds, then for the simultaneous update algorithm we have* [4]*:*

$$P_{\text{sim}} \geq \frac{1}{1 - \bar{\alpha}_\psi} D^\star.$$

---

[3] If $\psi(0) \neq 0$, then the definition of α_ψ should be modified to $\alpha_\psi(u) = \sup\{c \mid \langle y, u \rangle \geq (c+1)(\psi(u) - \psi(0)), y \in \partial\psi(u)\}$.

[4] If $\psi(0) \neq 0$, then the conclusion becomes $P_{\text{sim}} - \psi(0) \geq \frac{1}{1-\alpha_\psi}(D^\star - \psi(0))$.

where D^* is the dual optimal objective and

$$\bar{\alpha}_\psi = \inf\{\alpha_\psi(u) \,|\, u \in K\}. \tag{3.10}$$

For the sequential update algorithm,

$$P_{\text{seq}} \geq \frac{1}{1-\bar{\alpha}_\psi}(D^\star + \sum_{t=1}^{m} \langle A_t \hat{x}_t, \hat{y}_{t+1} - \hat{y}_t \rangle).$$

Further, if ψ is differentiable with gradient Lipschitz continuity parameter $\frac{1}{\mu}$ with respect to $\|\cdot\|$,

$$P_{\text{seq}} \geq \frac{1}{1-\bar{\alpha}_\psi}(D^\star - \sum_{t=1}^{m} \frac{1}{2\mu}\|A_t \hat{x}_t\|^2). \tag{3.11}$$

Proof. We first show that Assumption 2 implies that

$$\begin{aligned} \sum_{t=1}^{m} \sigma_t(A_t^T \tilde{y}_t) &\geq \sum_{t=1}^{m} \sigma_t(A_t^T \tilde{y}_{m+1}), \\ \sum_{t=1}^{m} \sigma_t(A_t \hat{y}_t) &\geq \sum_{t=1}^{m} \sigma_t(A_t \hat{y}_{m+1}), \end{aligned} \tag{3.12}$$

To do so, we prove that for all t, $\sigma_t(A_t^T \tilde{y}_t) \geq \sigma_t(A_t^T \tilde{y}_{m+1})$. The proof for $\sigma_t(A_t^T \hat{y}_t) \geq \sigma_t(A_t^T \hat{y}_{m+1})$ follows the same steps.

For any t, we have $\sum_{s=1}^{t} A_s \tilde{x}_s \leq_K \sum_{s=1}^{m} A_s \tilde{x}_s$ since $A_s F_s \subset K$ for all s. Since $\tilde{y}_t \in \partial\psi(\sum_{s=1}^{t} A_s \tilde{x}_s)$ and $\tilde{y}_{m+1}$ was chosen to be the minimum element in $\partial\psi(\sum_{s=1}^{m} A_s \tilde{x}_s)$ with respect to $\leq_{K^*}$, by the Assumption 2, we have

$$\tilde{y}_t \geq_{K^*} \tilde{y}_{m+1}.$$

Since $A_t x \in K$ for all $x \in F_t$, we get $\langle A_t x, y_t \rangle \geq \langle A_t x, y_m \rangle$. Therefore, $\sigma_t(A_t^T \tilde{y}_t) \geq \sigma_t(A_t^T \tilde{y}_{m+1})$. Now by (3.12)

$$D_{\text{sim}} = \sum_{t=1}^{m} \sigma_t(A_t^T \tilde{y}_t) - \psi^*(\tilde{y}_{m+1}) \geq \sum_{t=1}^{m} \sigma_t(A_t^T \tilde{y}_{m+1}) - \psi^*(\tilde{y}_{m+1}) \geq D^\star.$$

Similarly, $D_{\text{seq}} \geq D^\star$. Lemma 4 and definition of $\bar{\alpha}_\psi$ give the desired result. □

3.2.2 Competitive ratio for non-monotone ψ

If ψ is not non-decreasing with respect to K, the algorithms are not guaranteed to increase the objective at each step; therefore, P_{seq} and P_{sim} are not guaranteed to be non-negative. However, if we add the assumption that $0 \in F_t$ for all t, then Algorithm 3 will not decrease the objective at any step,

$$\psi(\sum_{s=1}^{t} A_s\tilde{x}_s) - \psi(\sum_{s=1}^{t-1} A_s\tilde{x}_s) \geq \langle \tilde{x}_t, A_t^T\tilde{y}_t\rangle = \max_{x\in F_t}\langle x, A_t^T\tilde{y}_t\rangle \geq 0, \tag{3.13}$$

which follows from concavity of ψ.

In other words, the algorithm simply assigns 0 to x_t if any other feasible point in F_t reduces the objective value. Define

$$\tilde{u} = \sum_{t=1}^{m} A_t\tilde{x}_t,$$

and note that under the assumption $0 \in F_t$, we have $\psi(\tilde{u}) \geq 0$. Further, we have:

$$P_{\text{sim}} \geq \frac{1}{1-\alpha_\psi(\tilde{u})} D^\star. \tag{3.14}$$

We provide examples of non-monotone ψ and their competitive ratio analysis in this section and section 3.3.1. Given appropriate conditions on F_t and A_t, in order to find a lower bound on the competitive ratio independent of $\tilde{u}$, we only need to lower bound α_ψ over a subset of K. Note that when ψ is not non-decreasing, then there exists a supergradient that is not in K^*. Therefore, $\alpha_\psi(\tilde{u})$ for general ψ can be less than -1. In this case, the lower bound for the competitive ratio of Algorithm (3) is less than $\frac{1}{2}$.

We now consider examples of ψ that satisfy Assumption 2 and derive lower bound on α_ψ for those examples.

3.2.3 Examples on the positive orthant.

Let $K = \mathbb{R}^n_+$ and note that $K^* = K$. To simplify the notation we use $\leq$ instead of $\leq_{\mathbb{R}^n_+}$. When ψ has continuous partial second derivatives with respect to all the variables, Assumption 2 is

equivalent to the Hessian being element-wise non-positive over $\mathbb{R}^n_+$. Assumption 2 is closely related to submodularity. In fact, on the positive orthant this assumption is sufficient for submodularity of ψ on the lattice defined by $\leq$. However, this assumption is not necessary for submodularity. When ψ has continuous partial second derivatives with respect to all the variables, the necessary and sufficient condition for submodularity only requires the off-diagonal elements of the Hessian to be non-positive [Lor53]. In Section 3.5, we delve deeper into connections with submodularity.

If ψ is separable, i.e.,

$$\psi(u) = \sum_{i=1}^{n} \psi_i(u_i), \tag{3.15}$$

Assumption 2 is satisfied since by concavity for each ψ_i we have $\partial\psi_i(u_i) \leq \partial\psi_i(v_i)$ when $u_i < v_i$. If ψ satisfies the Assumption 2, then so does $\psi(Au)$, where A is an element-wise non-negative matrix.

Adwords In the basic adwords problem, for all t, $F_t = \{x \in \mathbb{R}^l_+ \mid \mathbf{1}^T x \leq 1\}$, A_t is a diagonal matrix with non-negative entries, and

$$\psi(u) = \sum_{i=1}^{n} u_i - \sum_{i=1}^{n} (u_i - 1)_+, \tag{3.16}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$. In this problem, $\psi^*(y) = \mathbf{1}^T(y - \mathbf{1})$. Since $0 \in \partial\psi(\mathbf{1})$ we have $\alpha_\psi = -1$ by (3.9); therefore, the competitive ratio of algorithm 3 is $\frac{1}{2}$. Let $r = \max_{t,i,j} A_{t,ij}$, then we have $\sum_{t=1}^{m} \langle A_t \hat{x}_t, \hat{y}_{t+1} - \hat{y}_t \rangle \leq nr$. Therefore, the competitive ratio of algorithm 6 goes to $\frac{1}{2}$ as r (bid to budget ratio) goes to zero. In adwords with concave returns studied in [DJ12], A_t is diagonal for all t and ψ is separable [5].

[5]Note that in this case one can remove the assumption that $\partial\psi_i \subset \mathbb{R}_+$ since if $\tilde{y}_{t,i} = 0$ for some t and i, then $\tilde{x}_{s,i} = 0$ for all $s \geq t$.

Online linear program and non-separable budget Recall that online LP is given by

$$\begin{aligned} \text{maximize} \quad & \sum_{t=1}^{m} c_t^T x_t + I_{\{\cdot \leq \mathbf{1}\}} \left(\sum_{t=1}^{m} B_t x_t \right) \\ & x_t \in F_t, \quad \forall t \in [m]. \end{aligned}$$

where $F_t = \{x \in \mathbf{R}_+^k \mid \mathbf{1}^T x \leq 1\}$ is the simplex. If a lower bound on the optimal dual variable $\min_i y_i^* \geq -l$ is given, then the LP can be written in the exact penalty form [Ber75]:

$$\text{maximize}_{x_t \in F_t} \sum_{t=1}^{m} c_t^T x_t + G\left(\sum_{t=1}^{m} B_t x_t \right). \tag{3.17}$$

where G is an l-Lipschitz continuous function with respect to l_1 norm given by

$$G(u) = -l \sum_{i=1}^{n} (u_i - 1)_+.$$

$-l$ is a lower bound on dual variable if

$$l > \max \left\{ \frac{c_{t,j}}{B_{t,ij}} \mid B_{t,ij} > 0,\ j \in [k],\ i \in [n] \right\}. \tag{3.18}$$

To prove this, by way of contradiction, we assume that $y_i^* \leq -l$ for some $i \in [n]$. Now by the definition of l, we have $c_{t,j} + (B_t^T y^*)_j < 0$ for all j such that $B_{t,ij} > 0$. On the other hand by the optimality condition (3.4), $c_t + B_t^T y^* \in N_{F_t}(x_t^*)$ where $N_{F_t}(x_t^*)$ is the normal cone of the simplex at x_t^*. Therefore, we should have $x_{t,j}^* = 0$ for all j such that $B_{t,ij} > 0$. This results in $(B_t x_t^*)_i = 0$ which yield $(\sum_{t=1}^m B_t x_t^*)_i = 0$. This means that the corresponding variable $y_i^* = 0$ which is a contradiction. With this choice of l, Algorithm 3 always maintains a feasible solution for the original problem when applied to (3.17). This follows from the fact that if $\tilde{y}_{t,i} = -l$ for some t and i, then $(B_t \tilde{x}_t)_i = 0$. Note that in the adwords problem $B_t = \mathbf{diag}(c_t)$ and $l = 1$.

For any $p \geq 1$ let $\mathcal{B}_p$ be the l_p-norm ball. In order to provide examples of non-separable G, we rewrite the function $\sum_{i=1}^n (u_i - 1)_+$ using the distance from $\mathcal{B}_\infty$. For any set $C \subset \mathbb{R}_+^n$, let $d_1(u, C) = \inf_{\bar{u} \in C} \|u - \bar{u}\|_1$ and note that $d_1(u, C)$ is 1-Lipschitz continuous with respect

to l_1 norm. We have:

$$\sum_{i=1}^{n}(1-u_i)_+ = d_1(u, \mathcal{B}_\infty). \tag{3.19}$$

Consider a problem with constraint $\|\sum_{t=1}^{m} B_t x_t\|_p \leq 1$. Given the bound on dual variable in (3.18), this problem can be equivalently written in the form of (3.17) with the exact penalty [Bur91, Theorem 5.5]

$$G(u) = -ld_1(u, \mathcal{B}_p).$$

When $p = \infty$, we get back (3.19).

For $p \geq 1$, although not separable, the function $G(u) = -d_1(u, \mathcal{B}_p)$ satisfies Assumption 2. For $\mathcal{B}_1$, that follows from the fact that $d_1(u, \mathcal{B}_1) = (\mathbf{1}^T u - 1)_+$; therefore, $\partial d_1(u, \mathcal{B}_1) = \mathbf{1}\partial f(\mathbf{1}^T u)$, where $f(x) = (x-1)_+$. The proof for $p > 1$ is given in Appendix C.3.1.

When ψ is given by 3.17 with $G(u) = -ld_1(u, \mathcal{B}_p)$, we have $\alpha_\psi(\tilde{u}) \geq -\frac{l}{\theta}$, where $\theta = \min_t \min_{x \in F_t} \frac{c_t^T x}{\mathbf{1}^T B_t x}$. The derivation of this bound is also given in Appendix C.3.1.

3.2.4 Examples on the positive semidefinite cone.

Let $K = S_+^n$ and note that $K^* = K$. An interesting example that satisfies Assumption 2 is $\psi(U) = \mathbf{tr}U^p$ with $p \in (0,1)$, where $\nabla\psi(U) = pU^{p-1}$ and $\bar{\alpha}_\psi = p-1$. This objective function is used in pth mean optimal experiment design.

Another example is $\psi(U) = \log\det(U + A_0)$, where $\nabla\psi(U) = (U+A_0)^{-1}$ and $\bar{\alpha}_\psi = -1$ since

$$\frac{\langle \nabla\psi(U), U\rangle}{\psi(U) - \psi(0)} = \frac{n - \mathbf{tr}(A_0+U)^{-1}A_0}{\log\det((A_0+U)A_0^{-1})} \to 0 \text{ as } \mathbf{tr}U \to \infty.$$

Maximizing $\log\det$ arises in several offline applications including D-optimal experiment design [Puk93], maximizing the Kirchhoff complexity of a graph [ZP08], Optimal sensor selection [JB09, SBV10]. An example of applications for maximizing the logdet of a projected Laplacian, as a function of graph edge weights, appears in connectivity control of mobile networks [ZP08].

For a differentiable trace function, Assumption 2 can be expressed in an equivalent, more explicit form. This is a consequence of *Löwner's theorem* for matrix monotone functions [Han13], an important result in matrix analysis. From Theorem 4.9 in [Han13] it follows that, if $H(U) = \sum_{i=1}^{n} h(\lambda_i(U))$ then H satisfies Assumption 2 for all n, *if and only if* there exists a positive measure μ supported on $[0, 1]$ such that

$$h(u) = \int_0^u y(u')\, du' \quad \text{where} \quad y(u) = \int_0^1 \frac{1}{u\lambda + (1-\lambda)}\, d\mu(\lambda). \tag{3.20}$$

This alternative description allows us to explicitly impose Assumption 2 in the optimization problem (3.49) for smoothing trace functions.

We derive the competitive ratio of the greedy algorithm with smoothing for online experiment design and Kirchhoff complexity maximization of a graph in section 3.3.

3.3 Smoothing of ψ for improved competitive ratio

The technique of "smoothing" an objective function, or equivalently adding a strongly convex regularization term to its conjugate function, has been used in several areas. In convex optimization, smoothing via conjugate function has been applied for improving the ill-conditioned objective functions [Ber14, Pol79]. A general version of this is due to Nesterov [Nes05], and has led to faster convergence rates of first order methods for non-smooth problems. In this section, we study how replacing ψ with a appropriately smoothed function ψ_S helps improve the performance of the two algorithms discussed in section 3.1, and show that it provides optimal competitive ratio for two of the problems mentioned in section 3.2, adwords and online LP. We then show how to maximize the competitive ratio of Algorithm 3 for a separable ψ and compute the optimal smoothing by solving a convex optimization problem. This allows us to *design* the most effective smoothing customized for a given ψ: we maximize the bound on the competitive ratio over the set of smooth functions (see subsection 3.3.2 for details).

Let ψ_S denote an upper semi-continuous concave function (a smoothed version of ψ), and suppose ψ_S satisfies Assumption 2. The algorithms we consider in this section are the same as Algorithms 6 and 3, but with ψ replacing ψ_S. Note that the competitive ratio is

computed with respect to the original problem, that is the offline primal and dual optimal values are still the same $P^\star$ and $D^\star$ as before. If we replace ψ with ψ_S in algorithms 6 and 3, the dual updates are modified to

$$\hat{y}_{t+1} \in \operatorname*{argmin}_y \langle \sum_{s=1}^{t} A_s \hat{x}_s, y\rangle - {\psi_S}^*(y),$$
$$(\tilde{x}_t, \tilde{y}_t) \in \arg\min_y \max_{x \in F_t} \langle y, A_t x + \sum_{s=1}^{t-1} A_s x_s\rangle - {\psi_S}^*(y).$$

From Lemma 4, we have that

$$D_{\text{sim}} \leq \psi_S \left(\sum_{t=1}^{m} A_t \tilde{x}_t \right) - \psi^*(\tilde{y}_{m+1}) \tag{3.21}$$
$$D_{\text{seq}} \leq \psi_S \left(\sum_{t=1}^{m} A_t \hat{x}_t \right) - \psi^*(\hat{y}_{m+1}) - \sum_{t=1}^{m} \langle A_t \hat{x}_t, \hat{y}_{t+1} - \hat{y}_t\rangle. \tag{3.22}$$

Similar to our assumption on $\psi(0)$, to simplify the notation, by replacing $\psi_S(\cdot)$ with $\psi_S(\cdot) - \psi_S(0)$, we assume $\psi_S(0) = 0$. Define

$$\alpha_{\psi,\psi_S}(u) = \sup\{c \,|\psi^*(y) \geq \psi_S(u) + (c-1)\psi(u), \forall y \in \partial\psi_S(u)\},$$

and

$$\bar{\alpha}_{\psi,\psi_S} = \inf\{\alpha_{\psi,\psi_S}(u) \mid u \in K\}.$$

Now the conclusion of Theorem 5 holds with $\bar{\alpha}_\psi$ replaced by $\bar{\alpha}_{\psi,\psi_S}$.

Theorem 6. *If ψ_S satisfies Assumption 2, then for the simultaneous update algorithm we have* [6]*:*

$$P_{\text{sim}} \geq \frac{1}{1 - \bar{\alpha}_{\psi,\psi_S}} D^\star.$$

Similarly, when ψ_S is non-monotone, inequality (3.14) holds with α_ψ replaced by α_{ψ,ψ_S}.

The next theorem provides a lower bound on the competitive ratio for the sequential algorithm applied to the smooth function ψ_S.

[6] If $\psi(0) \neq 0$, then the conclusion becomes $P_{\text{sim}} - \psi(0) \geq \frac{1}{1-\alpha_\psi}(D^\star - \psi(0))$.

Theorem 7. *Suppose ψ_S is differentiable on an open set containing K and satisfies Assumption 2. In addition suppose there exists $c \in K$ is such that $A_t F_t \leq_K c$ for all t, then*

$$P_{\text{seq}} \geq \frac{1}{1 - \kappa_{\psi,\psi_S}} D^\star,$$

where κ is given by

$$\kappa_{\psi,\psi_S} = \sup\{c \mid \psi^*(\nabla\psi(u)) + \langle c, \nabla\psi_S(u) - \nabla\psi_S(0)\rangle \geq \psi_S(u) + (c-1)\psi(u), u \in K\} \quad (3.23)$$

Proof. Since ψ_S satisfies Assumption 2, we have $\hat{y}_{t+1} \leq_{K^*} \hat{y}_t$. Therefore, we can write:

$$\begin{aligned}\sum_{t=1}^{m} \langle A_t\hat{x}_t, \hat{y}_t - \hat{y}_{t+1}\rangle &\leq \sum_{t=1}^{m} \langle c, \hat{y}_t - \hat{y}_{t+1}\rangle \\ &= \langle c, \hat{y}_0 - \hat{y}_{m+1}\rangle \end{aligned} \quad (3.24)$$

Now by combining 3.22 with 3.24, we get

$$D_{\text{seq}} \leq \psi_S\left(\sum_{t=1}^{m} A_t\hat{x}_t\right) + \langle c, \nabla\psi_S(0) - \nabla\psi_S\left(\sum_{t=1}^{m} A_t\hat{x}_t\right)\rangle.$$

The conclusion of the theorem follows from the definition of κ_{ψ,ψ_S} and the fact that $\hat{D} \geq D^\star$.

□

3.3.1 Smoothing via the Conjugate Function

We first consider Nesterov smoothing [Nes05], and apply it to examples on non-negative orthant. Given a proper upper semi-continuous concave function $\phi : \mathbb{R}^n \mapsto \mathbb{R} \cup \{-\infty\}$, let

$$\psi_S = (\psi^* + \phi^*)^*.$$

Note that ψ_S is the hypo-sum (sup-convolution) of ψ and ϕ.

$$\psi_S = \psi\square\phi(u) = \sup_v \psi(v) + \phi(u - v).$$

This is called hypo-sum of ψ and ϕ since the hypo-graph of ψ_S is the Minkowski sum of hypo-graphs of ψ and ϕ.

If ψ and ϕ are separable, then $\psi\square\phi$ satisfies Assumption 2 for $K = \mathbb{R}^n_+$. Here we provide example of Nesterov smoothing for functions on non-negative orthant.

Adwords and Online LP: Consider the problem (3.17) with $G(u) = -l\sum_{i=1}^{n}(u_i - 1)_+$. For this problem we smooth G with:

$$\phi^*(y) = \frac{1}{\gamma}\sum_{i=1}^{m}\left((y_i - \frac{\theta}{e-1})\log(1 - \frac{e-1}{\theta}y_i)\right) - (1 + 1/\gamma)\mathbf{1}^T y, \tag{3.25}$$

where $\gamma = \log(1 + \frac{l(e-1)}{\theta})$,

$$\theta = \min_t \min_{x \in F_t} \frac{c_t^T x}{\mathbf{1}^T B_t x}$$

and l is defined as in (3.18). In this case, we have;

$$\alpha_{\psi,\psi\square\phi}(\tilde{u}) \geq 1 - (1 + \frac{1}{e-1})\gamma.$$

(see Appendix **??** for the derivation). This gives the competitive ratio of $\frac{1}{\gamma}(1 - 1/e)$. In the case of adwords where $\theta = l$, this yield the optimal competitive ratio of $1 - e^{-1}$ and the smoothed function coincides with the one derived in the previous paragraph. For a general LP, this approach provides the optimal competitive ratio, which is known to be $O(\gamma^{-1})$ [BN09].

3.3.2 Computing optimal smoothing for separable functions on $\mathbb{R}^n_+$

We now tackle the problem of finding the optimal smoothing for separable functions on the positive orthant, which as we show in an example at the end of this section is not necessarily given by Nesterov smoothing. Given a separable monotone $\psi(u) = \sum_{i=1}^{n}\psi_i(u_i)$ and $\psi_S(u) = \sum_{i=1}^{n}\psi_{S_i}(u_i)$ on $\mathbb{R}^n_+$ we have that $\bar{\alpha}_{\psi,\psi_S} \geq \min_i \bar{\alpha}_{\psi_i,\psi_{S_i}}$.

To simplify the notation, drop the index i and consider $\psi : \mathbb{R}_+ \mapsto \mathbb{R}$. We formulate the problem of finding ψ_S to maximize α_{ψ,ψ_S} as an optimization problem. In section 3.4 we discuss the relation between this optimization method and the optimal algorithm presented in [DJ12]. We set $\psi_M(u) = \int_0^u y(s)ds$ with y a continuous function, and state the infinite

dimensional convex optimization problem with y as a variable,

$$\begin{aligned} \text{minimize} \quad & \beta \\ \text{subject to} \quad & \int_0^u y(s)ds - \psi^*(y(u)) \leq \beta\psi(u), \qquad \forall u \in [0,\infty) \\ & y \in C[0,\infty). \end{aligned} \tag{3.26}$$

where $\beta = 1 - \bar{\alpha}_{\psi,\psi_S}$ (theorem 5 describes the dependence of the competitive ratios on this parameter). Note that we have not imposed any condition on y to be non-increasing (i.e., the corresponding ψ_S to be concave). The next lemma establishes that every feasible solution to the problem (3.49) can be turned into a non-increasing solution.

Lemma 5. *Let (y,β) be a feasible solution for problem* (3.49) *and define $\bar{y}(u) = \inf_{s\leq u} y(s)$. Then $(\bar{y},\beta)$ is also a feasible solution for problem* (3.49).

In particular if (y,β) is an optimal solution, then so is $(\bar{y},\beta)$. The proof is given in the supplement. Revisiting the adwords problem, we observe that the optimal solution is given by $y(u) = \left(\frac{e-\exp(u)}{e-1}\right)_+$, which is the derivative of the smooth function we derived using Nesterov smoothing in section 3.3.1. The optimality of this y can be established by providing a dual certificate, a measure ν corresponding to the inequality constraint, that together with y satisfies the optimality condition. If we set $d\nu = f(u)\,du$ with $f(u) = \exp(1-u)/(e-1)$, the optimality conditions are satisfied with $\beta = (1-1/e)^{-1}$.

$$\begin{aligned} &\int_{u=0}^{\infty} f(u)\psi(u)\,du = 1, \quad f \geq 0, \\ &\int_{s=u}^{\infty} f(s)\,ds \in f(u)\partial\psi^*(y(u)), \qquad \forall u \geq 0, \\ &\int_0^u y(s)ds - \psi^*(y(u)) \leq \beta\psi(u), \qquad \forall u \geq 0, \\ &f(u)(\int_0^u y(s)ds - \psi^*(y(u)) - \beta\psi(u)) = 0, \qquad \forall u \geq 0. \end{aligned}$$

Also note that if ψ plateaus (e.g., as in the adwords objective), then one can replace problem (3.49) with a problem over a finite horizon.

Theorem 8. *Suppose $\psi(u) = c$ on $[u', \infty)$. Then problem* (3.49) *is equivalent to the following problem,*

$$\begin{aligned} \text{minimize} \quad & \beta \\ \text{subject to} \quad & \int_0^u y(s)ds - \psi^*(y(u)) \leq \beta\psi(u), \qquad \forall u \in [0, u'] \\ & y(u') = 0, \quad y \in C[0, u']. \end{aligned} \tag{3.27}$$

So for a function ψ with a plateau, one can discretize problem (3.27) to get a finite dimensional problem,

$$\begin{aligned} \text{minimize} \quad & \beta \\ \text{subject to} \quad & h\sum_{s=1}^{t} y[s] - \psi^*(y[t]) \leq \beta\psi(ht), \qquad \forall t \in [d] \\ & y[d] = 0, \end{aligned} \tag{3.28}$$

where $h = u'/d$ is the discretization step. Figure 3.2A shows the optimal smoothing for the piecewise linear function $\psi(u) = \min(.75,\ u,\ .5u + .25)$ by solving problem (3.28). We point out that the optimal smoothing for this function is *not* given by Nesterov's smoothing (even though the optimal smoothing can be derived by Nesterov's smoothing for a piecewise linear function with only two pieces, like the adwords cost function). Figure 3.2D shows the difference between the conjugate of the optimal smoothing function and ψ^* for the piecewise linear function, which we can see is not concave.

In cases where a bound $u_{\max}$ on $\sum_{t=1}^{m} A_t F_t$ is known, we can restrict t to $[0, u_{\max}]$ and discretize problem (3.49) over this interval. However, the conclusion of Lemma 5 does not hold for a finite horizon and we need to impose additional linear constraints $y[t] \leq y[t-1]$ to ensure the monotonicity of y. We find the optimal smoothing for two examples of this kind: $\psi(u) = \log(1+u)$ over $[0, 100]$ (Figure 3.2B), and $\psi(u) = \sqrt{u}$ over $[0, 100]$ (Figure 3.2C). In Figure 3.2E, we show the competitive ratio achieved with the optimal smoothing

of $\psi(u) = \log(1+u)$ over $[0, u_{\max}]$ as a function of $u_{\max}$. Figure 3.2F depicts this quantity for $\psi(u) = \sqrt{u}$.

Smoothing for the sequential algorithm.

We provide a lower-bound on the competitive ratio of the sequential algorithm (Algorithm 6). Based on this competitive ratio bound we modify Problem (3.49) for designing the smoothing function

Based on the result of the previous theorem we can modify the optimization problem set up in Section 3.3.2 for separable functions on $\mathbf{R}^n_+$ to maximize the lower bound on the competitive ratio of the sequential algorithm. Note that in this case we have $\kappa_{\psi,\psi_S} \leq \max_i \kappa_{\psi_i,\psi_{S_i}}$. Similar to the previous section to simplify the notation we drop the index i and assume ψ is a function of a scalar variable. The optimization problem for finding ψ_S that minimizes $\kappa_{\psi,\psi_S} - \bar{\alpha}_{\psi,\psi_S}$ is as follows:

$$\begin{aligned} \text{minimize} \quad & \beta \\ \text{subject to} \quad & \int_0^u y(s)ds + c(\psi'(0) - y(u)) - \psi^*(y(u)) \leq \beta\psi(u), \qquad \forall u \in [0,\infty) \\ & y \in C[0,\infty). \end{aligned} \tag{3.29}$$

In the case of Adwords, the optimal solution is given by

$$\beta = \frac{1}{1-\exp(\frac{-1}{c+1})}, \quad y(u) = \beta\left(1 - \exp\left(\frac{u-1}{1+c}\right)\right)_+,$$

which gives a competitive ratio of $1 - \exp\left(\frac{-1}{c+1}\right)$. In Figure 3.3B we have plotted the competitive ratio achieved by solving problem 3.54 for $\psi(u) = \log\det(1+u)$ with $u_{\max} = 100$ as a function of c. Figure 3.3A shows the competitive ratio as a function of c for the piecewise linear function $\psi(u) = \min(.75,\ u,\ .5u + .25)$.

3.3.3 Computing optimal smoothing for trace function on the PSD cone

We address the problem of designing smoothing for problem (3.2) where H is a trace functions over PSD cone. Similar to problem (3.17), by Clarke's exact penalty principle [Cla90], we

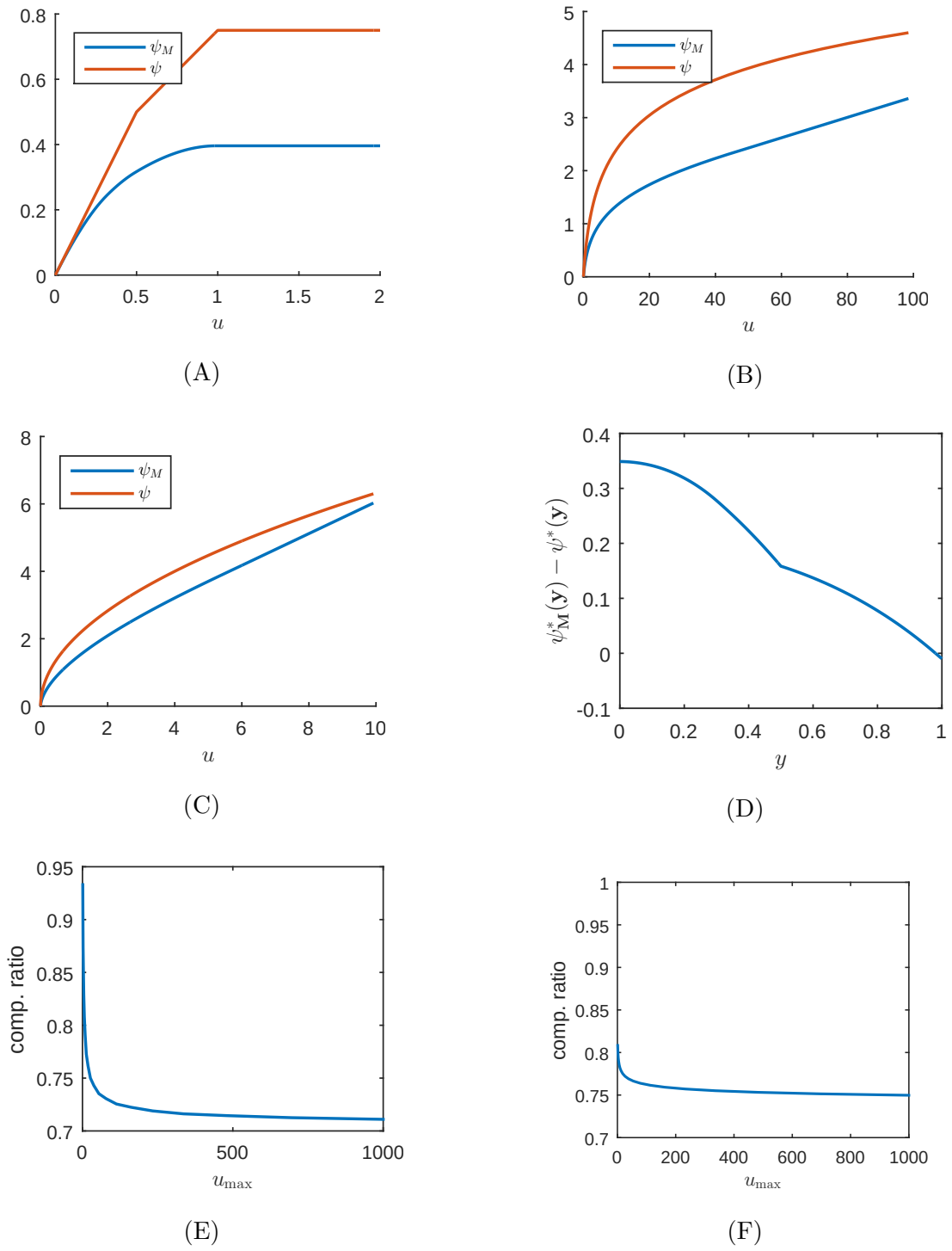
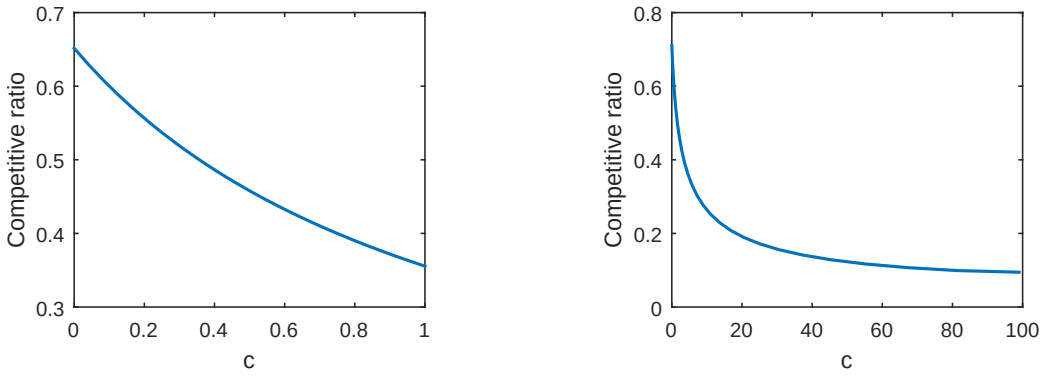


Figure 3.2: Optimal smoothing for $\psi(u) = \min(.75,\ u,\ .5u+.25)$ (**A**), $\psi(u) = \log(1+u)$ over $[0, 100]$ (**B**), and $\psi(u) = \sqrt{u}$ over $[0, 10]$ (**C**). The competitive ratio achieved by the optimal smoothing as a function of $u_{\max}$ for $\psi(u) = \log(1+u)$ (**E**) and $\psi(u) = \sqrt{u}$ (**F**). $\psi_M^* - \psi^*$ for the piecewise linear function (**D**).



(A) (B)

Figure 3.3: The competitive ratio achieved by the optimal smoothing for the sequential algorithm as a function of c for $\psi(u) = \min(.75,\ u,\ .5u + .25)$ (a) and $\psi(u) = \log(1 + u)$ with $u_{\max} = 100$ (b).

can write (3.2), for a sufficiently large positive constant l, problem (3.1) can be equivalently written as

$$\begin{array}{ll} \text{maximize} & H\left(\sum_{t=1}^{m} A_t x_t\right) + G(\sum_{t=1}^{m} x_t), \\ \text{subject to} & x_t \in [0, c_t], \quad t = 1, 2, \ldots, m. \end{array} \tag{3.30}$$

where

$$G\left(u\right) = -l\left(u - b\right)_+ .$$

The following lemma gives a bound on the parameter l that ensures (3.1) and (3.30) are equivalent. The proof is given in Appendix C.5.

Lemma 6. *If* $l > h'\left(0\right) \max_t \mathbf{tr} A_t$ *, then the optimization problems* (3.1) *and* (3.30) *are equivalent.*

The correspondence between this problem and problem (3.1) can be made explicit by taking the cone to be $K = \mathbf{S}_+^n \times \mathbf{R}_+$, the objective to be $\psi(U, u) = H(U) + G(u)$, and the

constraints to be defined by $F_t = [0, c_t]$ for $t = 1, 2, \ldots, m$. The dual of (3.30) can be written using conjugate functions as

$$\underset{z,Y}{\text{minimize}} \quad \sum_{t=1}^{m} c_t \left(\langle A_t, Y\rangle - z\right)_+ - H^*\left(Y\right) - G^*\left(z\right). \tag{3.31}$$

Our goal is to design a concave differentiable trace function H_S and a concave differentiable function G_S (these functions are smooth versions of H and G) to maximize the our competitive ratio lower-bounds for Algorithms 6 and Algorithms 3 applied to $\psi_S(U, u) = H_S(U) + G_S(u)$ (see Algorithm (4) and Algorithm (5)).

Algorithm 4 Sequential Update applied to $H_S(U) + G_S(u)$

Initialize $\hat{z}_0 = G_S'\left(0\right), \hat{Y}_0 = \nabla H_S\left(0\right)$

for $i \leftarrow 1\ m$ **do**

Receive A_t, F_t

$$\hat{x}_t = \begin{cases} c_t, & \hat{z}_{t-1} + \langle A_t, \hat{Y}_{t-1}\rangle > 0 \\ 0, & \hat{z}_{t-1} + \langle A_t, \hat{Y}_{t-1}\rangle \leq 0 \end{cases}$$

$\hat{Y}_t = \nabla H_S\left(\sum_{s=1}^{t} A_s \hat{x}_s\right)$

$\hat{z}_t = G_S'\left(\sum_{s=1}^{t} \hat{x}_s\right)$

Algorithm 5 Simultaneous Update applied to $H_S(U) + G_S(u)$

for $i \leftarrow 1\ m$ **do**

Receive A_t, c_t

$\left(\tilde{z}_t, \tilde{Y}_t, \tilde{x}_t\right) \in$

$\arg\min_{z,Y} \max_{x \in [0, c_t]} \langle Y, A_t x + \sum_{s=1}^{t-1} A_t \tilde{x}_s\rangle$

$+ \langle z, x + \sum_{s=1}^{t-1} \tilde{x}_s\rangle - H_S^*\left(Y\right) - G_S^*\left(z\right)$

To design H_S and G_S, we need access to l (an uper bound on $h'\left(0\right) \max_t \mathbf{tr} A_t$) and a lower bound on the trace of all A_t's:

$$\theta \leq \min_t \mathbf{tr} A_t. \tag{3.32}$$

Given such a bound we have

$$H\left(\sum_{t=1}^{m} A_t x_t\right) \geq h\left(\theta \sum_{t=1}^{m} x_t\right). \tag{3.33}$$

To see this, let $U = \sum_{t=1}^{m} A_t x_t$ and $u = \sum_{t=1}^{m} x_t$. Note that by the definition of θ, we have $\sum_{j=1}^{n} \lambda_j(U) \geq \theta u$. We can write:

$$\begin{aligned} H(U) &= \sum_{i=1}^{n} h(\lambda_i(U)) \geq \sum_{i=1}^{n} \frac{\lambda_i(U)}{\sum_{j=1}^{n} \lambda_j(U)} h\left(\sum_{j=1}^{n} \lambda_j(U)\right) + \left(1 - \frac{\lambda_i(U)}{\sum_{j=1}^{n} \lambda_j(U)}\right) h(0) \\ &= \frac{\sum_{i=1}^{n} \lambda_i(U)}{\sum_{j=1}^{n} \lambda_j(U)} h\left(\sum_{j=1}^{n} \lambda_j(U)\right) \geq h(\theta u), \end{aligned}$$

where we used the concavity of h in the first line and its monotonicity alongside the fact that $h(0) = 0$ in the second line. Without loss of generality, by changing variables x_t to θx_t, and replacing b with θb and A_t with $\frac{1}{\theta} A_t$, we can assume that $\theta = 1$. We first drive competitive ratio results that allow us to search for functions h_S and G_S that maximizes the competitive ratio bounds.

Design of smoothing function for the simultaneous algorithm.

We first present a theorem that lower bounds the competitive ratio for the simultaneous algorithm.

Theorem 9. *Let* $u_{\max} \geq \lambda_{\max}\left(\sum_{t=1}^{m} A_t \tilde{x}_t\right)$. [7] *If* H_S *satisfies assumption 2 on* S_+^n *and there exist* $\gamma \geq 0$ *and* $\beta \geq 0$ *such that*

$$\gamma h_S(u) \leq h^*\left(h_S'(u)\right) + \beta h(u), \quad \text{for all } u \in [0,\ u_{\max}] \tag{3.34}$$

$$\gamma G_S(u) \leq G^*\left(G_S'(u)\right) + \gamma h(u)/(e-1), \quad \text{for all } u \in [0,\ b], \tag{3.35}$$

then

$$H\left(\sum_{t=1}^{m} A_t \tilde{x}_t\right) \geq \frac{1}{\gamma/(e-1) + \beta} D^\star \qquad \text{and} \qquad \sum_{t=1}^{m} \tilde{x}_t \leq b', \tag{3.36}$$

[7] We can always take, for instance, $u_{\max} = b\rho_2$ to satisfy this bound, but for specific examples it may be possible to choose a smaller value of $u_{\max}$.

where $b' = \inf\{u \mid G'_S(u) < -h'(0)\max_t \mathbf{tr}A_t\}$.

In particular, if

$$G'_S(u) < -h'(0)\max_t \mathbf{tr}A_t, \quad \text{for all } u > b,, \tag{3.37}$$

then $\sum_{t=1}^m \tilde{x}_t \le b$ *and*

$$P_{\text{sim}} \ge \frac{1}{\gamma/(e-1)+\beta} D^\star.$$

Proof. Let $U = \sum_{t=1}^m A_t\tilde{x}_t$, $u = \sum_{t=1}^m \tilde{x}_t$, $Y = \nabla H_S(U)$, and $z = G'_S(u)$. By assumption, $\lambda_{\max}(U) \le u_{\max}$. First we show that $u \le b'$. If $\tilde{z}_t < -h'(0)\max_s \mathbf{tr}A_s$ for some t, then since

$$\langle A_t, \nabla H_S\left(\textstyle\sum_{s=1}^t A_s\tilde{x}_s\right)\rangle \le h'\left(\lambda_{\min}\left(\textstyle\sum_{s=1}^t A_s\tilde{x}_s\right)\right)\mathbf{tr}A_t \le h'(0)\,\mathbf{tr}A_t,$$

it follows that $\langle \tilde{Y}_t, A_t\rangle + \tilde{z}_t < 0$ which results in $\tilde{x}_t = 0$. Given that $G_S{}'(u) < -h'(0)\max_t \mathbf{tr}A_t$ when $u > b'$, we conclude $u \le b'$.

Now by the duality gap bound (Lemma 4) we have

$$H_S(U) + G_S(u) - D_{\text{sim}} \ge H^*(Y) + G^*(y). \tag{3.38}$$

Also by the primal allocation rule in Algorithm 3, we have $\tilde{x}_t\left(\tilde{z}_t + \langle A_t, \tilde{Y}_t\rangle\right) \ge 0$. Combining these two observations with the concavity of H_S and G_S, we get

$$H_S\big(\textstyle\sum_{s=1}^t A_t\tilde{x}_s\big) + G_S\big(\sum_{s=1}^t \tilde{x}_s\big) - H_S\big(\sum_{s=1}^{t-1} A_t\tilde{x}_s\big) - G_S\big(\sum_{s=1}^{t-1}\tilde{x}_s\big) \ge \tilde{x}_t\Big(\tilde{z}_t + \langle A_t, \tilde{Y}_t\rangle\Big) \ge 0.$$

By taking the sum over t and telescoping the sum we get

$$H_S\left(\textstyle\sum_{s=1}^m A_t\tilde{x}_s\right) + G_S\left(\sum_{s=1}^m \tilde{x}_s\right) \ge 0. \tag{3.39}$$

Now we can write

$$\begin{aligned}
H(U) - D_{\text{sim}} &\ge -H_S(U) - G_S(u) + H^*(Y) + G^*(z) + H(U) + G(u) && \text{By (3.38)}\\
&\ge -H_S(U) + (\gamma-1)G_S(u) + \big(1-\tfrac{\gamma}{e-1}\big)H(U) + H^*(Y) && \text{By (3.35) and (3.33)}\\
&\ge -H_S(U) + (1-\gamma)H_S(U) + \big(1-\tfrac{\gamma}{e-1}\big)H(U) + H^*(Y) && \text{By (3.39) and } \gamma \ge 1\\
&= -\gamma H_S(U) + \big(1-\tfrac{\gamma}{e-1}\big)H(U) + H^*(Y)\\
&\ge \big(1-\tfrac{\gamma}{e-1}-\beta\big)H(U) && \text{By (3.34).}
\end{aligned}$$

Similar to the proof of theorem 5, since H_S and G_S satisfy Assumption 2, we have $D_{\text{sim}} \geq D^\star$. This combined with the previous inequality proves the conclusion of the theorem. □

Our aim is to find functions h_S and G_S and positive scalars β and γ so as to maximize $1/(\beta + \gamma/(e-1))$ subject to the constraints (3.34), (3.35), (3.52) of Theorem 9, and the constraint that H_S satisfies Assumption 2. Rather than working with h_S and G_S, we use the variables $y = h_S'$ and $z = G_S'$. Then $h_S(u) = \int_0^u y(v)\,dv$ and $G_S(u) = \int_0^u z(v)\,dv$. As such, we aim to solve the optimization problem

$$\begin{aligned}
&\text{minimize}_{\beta,\gamma,y,z,\mu} && \beta + \gamma/(e-1) && (3.40)\\
&\text{subject to} && \gamma \int_0^u y(v)\,dv \leq h^*\left(y(u)\right) + \beta h\left(u\right), \quad \text{for all } u \in [0, u_{\max}] && (3.41)\\
& && \gamma \int_0^u z(v)\,dv \leq bz(u) + \gamma h\left(u\right)/(e-1), \quad \text{for all } u \in [0, b] && (3.42)\\
& && z\left(u\right) = -l, \quad \text{for all } u \geq b, \quad z \in C[0,\infty) && (3.43)\\
& && y(u) = \int_0^1 \frac{1}{u\lambda + (1-\lambda)}\,d\mu\left(\lambda\right) && (3.44)\\
& && \mu \text{ a positive measure supported on } [0,1]. && (3.45)
\end{aligned}$$

Here we have used the explicit form of G^* in (3.42). Observe that the constraint (2) is imposing the requirement that H_S satisfies Assumption 2 by making use of the representation given by (3.20). Also, observe that the only constraint that involves both γ and β is (3.41). Now since $y \geq 0$ by (3.44), the constraint (3.41) is looser for smaller γ. Therefore, we can break this optimization problem into two problems. We first optimize for γ and z, and then optimize for the remaining decision variables. It is straightforward to check that partially optimizing for γ and z can be achieved by solving

$$\begin{aligned}
&\text{minimize}_{\gamma,z\in C[0,\infty)} && \gamma && \\
&\text{subject to} && \textstyle\int_0^u z\left(v\right)dv \leq \frac{b}{\gamma} z\left(u\right) + h\left(u\right)/(e-1) \quad \text{for all } u \in [0,b] && (3.46)\\
& && z\left(u\right) < -l \quad \text{for all } u > b &&
\end{aligned}$$

While this problem is not convex, it is linear in z for fixed γ. Remarkably, this allows us to actually find an explicit solution for (3.46).

Proposition 3. *The optimal solution of* (3.46) *is given by*

$$z\left(u\right)=-\frac{\gamma}{b(e-1)}\exp\left(\frac{\gamma}{b}u\right)\int_{0}^{u}\exp\left(-\frac{\gamma}{b}v\right)h'\left(v\right)\,dv \tag{3.47}$$

where γ is the solution to the following equation,

$$\frac{\gamma}{b(e-1)}\exp\left(\gamma\right)\int_{0}^{b}\exp\left(-\frac{\gamma}{b}v\right)h'\left(v\right)\,dv=l. \tag{3.48}$$

Moreover, $\gamma\geq 1$.

Proof. To argue for the optimality of (z,γ) given by the above equations, we consider any feasible solution $(\bar{z},\bar{\lambda})$ to (3.46). Since $\int_0^u \bar{z}\left(v\right)dv \leq h\left(u\right)+\frac{b}{\bar{\gamma}}z\left(u\right)$ for all $u\in[0,b]$, by Gronwall's theorem (see [Dra03] Corollary 2), we get

$$\bar{z}(u)\geq-\frac{\bar{\gamma}}{b(e-1)}\exp\left(\frac{\bar{\gamma}u}{b}\right)\int_{v=0}^{u}\exp\left(-\frac{\bar{\gamma}}{b}v\right)h'(v)\,dv\quad\forall u\in[0,b].$$

Therefore, since $\bar{z}(b)\leq -l$, it follows that $-\frac{\bar{\gamma}}{b(e-1)}\exp\left(\bar{\gamma}\right)\int_{v=0}^{b}\exp\left(\frac{-\bar{\gamma}}{b}v\right)h'(v)\,dv\leq -l$. This yields $\bar{\gamma}\geq\gamma$, by monotonicity of the left-hand side of (3.48), which is established in Appendix C.5.1. To show that $\gamma\geq 1$, we can write:

$$\begin{aligned}l&=\frac{\gamma}{b(e-1)}\exp\left(\gamma\right)\int_{0}^{b}\exp\left(-\frac{\gamma}{b}v\right)h'\left(v\right)\,dv\\&\leq\frac{\gamma h'\left(0\right)}{b(e-1)}\exp\left(\gamma\right)\int_{0}^{b}\exp\left(-\frac{\gamma}{b}v\right)\,dv=\frac{h'\left(0\right)\left(\exp(\gamma)-1\right)}{e-1}\end{aligned}$$

Now by our assumption that $\theta=1$ and by the definition of l, we have $l\geq h'(0)$. This combined with previous inequality establishes $\gamma\geq 1$. □

Once the optimum γ is found, we can substitute it into (3.41) and minimize β. Note that if the optimum γ is less than 1, we can simply substitute 1 instead of γ in (3.41) and z still remains a feasible for (3.42) with γ replaced by 1. We can use (3.44) to express y in terms

of μ, eliminating y from the formulation. Doing this, we obtain:

$$\begin{aligned} &\text{minimize}_{\beta,\mu} \ \ \beta \\ &\text{subject to} \ \ \gamma \int_0^1 \frac{\log\left(\frac{u\lambda}{1-\lambda}+1\right)}{\lambda} d\mu\left(\lambda\right) - h^*\left(\int_0^1 \frac{d\mu(\lambda)}{u\lambda+(1-\lambda)}\right) \leq \beta h\left(u\right), \ \forall u \in [0, u_{\max}] \\ &\qquad \mu \text{ a positive measure supported on } [0,1]. \end{aligned} \tag{3.49}$$

We discuss how to solve these optimization problems numerically in Section 3.3.4.

Smoothing for the sequential algorithm

The next theorem is an analogue of Theorem 9 but for the sequential algorithm. Since Algorithm 3 uses the dual variable from the previous time step to assign the primal variable, the maximum length of each primal step, which is captured by two parameters (ρ_1 and ρ_2), plays a role in the competitive ratio.

Theorem 10. *Let $\rho_1 \geq \max_t c_t$, $\rho_2 \geq \max_t \lambda_{\max}\left(A_t\right)$, and $u_{\max} \geq \lambda_{\max}\left(\sum_{t=1}^m A_t \hat{x}_t\right)$. If H_S satisfies Assumption 2 on S_+^n, and If there exist $\gamma > 0$ and $\beta > 0$ such that*

$$\gamma\left[h_S\left(u\right) + \rho_1\rho_2\left(h_S'\left(0\right) - h_S'\left(u\right)\right)\right] \leq h^*\left(h_S'(u)\right) + \beta h\left(u\right), \quad \text{for all } u \in [0, u_{\max}] \tag{3.50}$$

$$\gamma\left[G_S\left(u\right) - \rho_1 G_S'\left(u\right)\right] \leq G^*\left(G_S'(u)\right) + \gamma h(u)/(e-1), \quad \text{for all } u \in [0, b-\rho_1] \tag{3.51}$$

then

$$H\left(\sum_{t=1}^m A_t\hat{x}_t\right) \geq \frac{1}{\gamma/(e-1)+\beta(\gamma)} D^\star \qquad \text{and} \qquad \sum_{t=1}^m \hat{x}_t \leq b',$$

where $b' = G_S'^{-1}(h'\left(0\right)\max_t \mathbf{tr} A_t) + \rho_1$. In particular, if

$$G_S'\left(u\right) < -h'\left(0\right)\max_t \mathbf{tr} A_t, \quad \text{for all } u > b - \rho_1,, \tag{3.52}$$

then $\sum_{t=1}^m \hat{x}_t \leq b$ and

$$P_{\text{seq}} \geq \frac{1}{\gamma/(e-1)+\beta} D^\star.$$

Proof. The proof is similar to that of Theorem 9 and is given in Appendix C.6. □

Design of smoothing function for the sequential algorithm.

The design of smoothing functions for the sequential algorithm is very similar to the design for the simultaneous algorithm, but now based on Theorem 10. The main difference is the presence of the parameters ρ_1 and ρ_2 from Theorem 10. To find G_S for the sequential algorithm we solve the following problem in the variable $z = G_S'$,

$$\underset{\gamma, z \in C[0,\infty)}{\text{minimize}}\, \gamma \quad \text{subject to} \quad \begin{cases} \int_0^u z(v)\, dv - \rho_1 z \leq \frac{h(u)}{e-1} + \frac{b}{\gamma} z(u) & \forall u \in [0, b-\rho_1] \\ z(u) < -l & \forall u > b - \rho_1 \end{cases} \tag{3.53}$$

Similar to Proposition 3, the optimal solution is given by

$$z(u) = -\frac{\gamma}{(b+\rho_1\gamma)(e-1)} \exp\left(\frac{\gamma}{b+\rho_1\gamma} u\right) \int_0^u \exp\left(-\frac{\gamma}{b+\rho_1\gamma} v\right) h'(v)\ dv,$$

where γ is the solution to the following equation,

$$\frac{\gamma}{(b+\rho_1\gamma)(e-1)} \exp\left(\frac{\gamma(b-\rho_1)}{b+\rho_1\gamma}\right) \int_0^{b-\rho_1} \exp\left(-\frac{\gamma}{b+\rho_1\gamma} v\right) h'(v)\ dv = l.$$

To find h_S for the sequential algorithm, the problem (3.49) is modified to:

$$\begin{aligned} \text{minimize}_{\beta,y,\mu} \quad & \beta \\ \text{subject to} \quad & \gamma \int_0^u y(s)\, ds + \rho_1\rho_2(y(0) - y(u)) - h^*(y(u)) \leq \beta h(u), \qquad \forall u \in [0, u_{\max}] \\ & y(u) = \int_0^1 \frac{1}{u\lambda + (1-\lambda)}\, d\mu(\lambda) \\ & \mu \text{ a positive measure supported on } [0,1]. \end{aligned} \tag{3.54}$$

Here we have explicitly kept y as a decision variable to simplify the description of the problem. It could be eliminated in the same way as in (3.49).

3.3.4 Numerical implementation and examples of smoothing design for trace functions

We briefly discuss the numerical implementation of the smoothing design problems for the simultaneous algorithm, introduced in Section 3.3.3. Similar ideas apply for the smoothing design problems for the sequential algorithm.

Given l, b, and h, we can solve (3.48) for γ by bisection-based one-dimensional root-finding. The formula (3.47) for z can then be computed up to desired accuracy using Gauss-Leguerre quadrature [Rei43]. We note that the algorithm does not require G_S itself, but only $G'_S(u) = z(u)$.

In practice, if an estimate of l is not available one can take γ in the definition of G'_S (3.47) as a design parameter. By Theorems 9 and 10 the amount of budget used by the algorithm (b') depends on G'_S which directly depends on γ.

To solve problem (3.49), we restrict μ to be an atomic measure supported on the $q+1$ points $\lambda_j = j/q \in [0,1]$ for $j = 0, 1, \ldots, q$. The decision variables are then β and $\mu_j := \mu(\lambda_j)$ for $j = 0, 1, \ldots, q$. Rather than imposing the constraint for all $u \in [0, u_{\max}]$ we impose it on a non-uniformly sampled subset. In particular, we sample u more densely where h has a larger local Lipschitz constant by choosing the discretization points to be $u_i = h^{-1}\left(i\, u_{\max}/d\right)$ for $i = 0, 1, \ldots, d$. This results in an approximation of (3.49) by the finite-dimensional convex optimization problem

$$\begin{aligned} &\underset{\beta,\mu\in\mathbf{R}_+^{q+1}}{\text{minimize}} \;\; \beta \qquad\qquad (3.55)\\ &\text{subject to} \;\; \gamma \sum_{j=0}^{q} \mu_j \frac{\log\left(\frac{u_i\lambda_j}{1-\lambda_j}+1\right)}{\lambda_j} - h^*\left(\sum_{j=0}^{q} \frac{\mu_j}{u_i\lambda_j + (1-\lambda_j)}\right) \le \beta h\left(u_i\right), \;\; \text{for } i = 0, 1, \ldots, d. \end{aligned}$$

The optimal $y = h'_S$, which is all that is needed for the algorithm, is $y(u) = \sum_{j=0}^{q} \frac{\mu_j}{u\lambda_j + (1-\lambda_j)}$.

Examples

We consider different examples of h and, for each, find h_S and G_S.

Example 1. Consider the linear function $h(u) = u$ which translates to $H(U) = \mathbf{tr}(U)$. This case allows us to sow that our approach recovers the Nesterov smoothing for online LP discussed in section 3.3.1. In this case, the problem reduces to a linear program with one

budget constraint:

$$\text{maximize} \quad \sum_{t=1}^{m} \mathbf{tr}(A_t)x_t \quad \text{subject to} \quad \begin{cases} \sum_{t=1}^{m} x_t \leq b \\ 0 \leq x_t \leq c_t, \quad \forall t = 1, 2, \ldots, m. \end{cases} \tag{3.56}$$

The parameter l should be chosen to satisfy $l > \max_t \mathbf{tr}(A_t)$. Proposition 3 tells us that $G'_S(u) = 1 - \exp\left(\frac{\gamma}{b}u\right)$ and $\gamma = \log(1+l)$ are optimal for G'_S and γ. Since $h(u) = u$ it follows that $\text{Dom}(h^*) = \{1\}$. From (3.41) we can then deduce that $h_S = h$ and that $\beta = \gamma$. Note that G_S can be derived as a smooth version of G using Nesterov smoothing with the shifted entropy function given in 3.25

$$G^*_S(z) = G^*(z) + \frac{b}{\gamma}((z - 1/(e-1))\log(1-(e-1)z) - (\gamma+1)z).$$

Example 2. We consider $h(u) = \log(u+1)$, i.e., $H(U) = \log\det(I+U)$. This example also covers $\log\det(A_0 + U)$ with A_0 invertible. This is because by replacing A_t with $A_0^{-1/2}A_tA_0^{-1/2}$ and using the logarithm product property, one can equivalently write the problem with $\log\det(I+U)$ as the objective. In this case, $h^*(y) = \log(y) - y + 1$. The optimal $G'_S = z$ is

$$z(u) = -\frac{\gamma}{b(e-1)} \exp\left(\frac{\gamma}{b}(u+1)\right)\left[\text{E}_1\left(\frac{\gamma}{b}\right) - \text{E}_1\left(\frac{\gamma}{b}(u+1)\right)\right]$$

where γ is chosen so that $z(b) = -l$ and E_i is the *exponential integral* defined as

$$\text{E}_\text{i}(u) = \int_1^\infty \frac{\exp(-uv)}{v^i}\, dv. \tag{3.57}$$

Figure 3.4 depicts an example G_S for $b = 10$ with two different values of γ and an example of h_S with $u_{\max} = 10$ and $\gamma = 4$. Figure 3.5E depicts γ as a function of l for different values of b. Notice that γ scales logarithmically with l. In Figure 3.5A, we plot the competitive ratio bound of Theorem 9 vs γ. As mentioned if l is not available to the algorithm designer, γ can be treated as a free parameter that should be chosen by the designer. In that case the budget violation depends on γ and in Figure 3.5C, we plot b'/b versus γ (as a free parameter), when $h(u) = \log(u+1)$.

We can observe from 3.5C that for a fixed γ the budget violation scales almost logarithmically in $h'(0)\max_t \mathbf{tr} A_t$. In fact a modified version of the worst case example in [BN09] shows that to achieve a constant competitive ratio, budget violation is $\Omega(b\log(h'(0)\max_t \mathbf{tr} A_t))$.

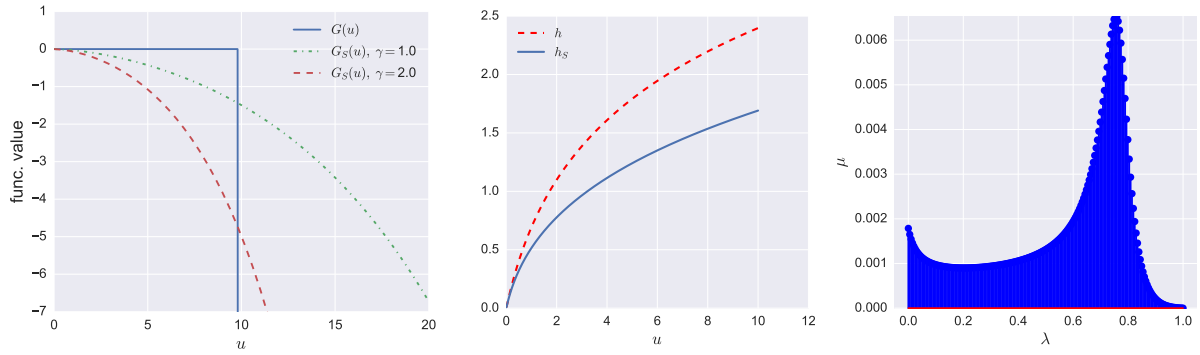


Figure 3.4: (a) Examples of G_S, and (b) h_S and the corresponding measure μ, when $h(u) = \log(u+1)$ and $\gamma = 2$.

For the sequential algorithm $G_S(u) = \int_0^u z(s)\,ds$ is given by:

$$z(u) = -\frac{\gamma}{(b+\rho_1\gamma)(e-1)} \exp\left(\frac{\gamma\,(u+1-\rho_1)}{b+\rho_1\gamma}\right)\left(\mathrm{E}_1\left(\frac{\gamma}{b+\rho_1\gamma}\right) - \mathrm{E}_1\left(\frac{\gamma}{b+\rho_1\gamma}\,(u+1-\rho_1)\right)\right)$$

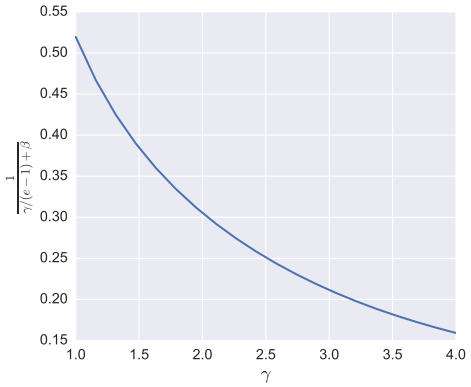
where γ satisfies:

$$\frac{\gamma}{(b+\rho_1\gamma)(e-1)} \exp\left(\frac{\gamma\,(b+1-\rho_1)}{b+\rho_1\gamma}\right)\left(\mathrm{E}_1\left(\frac{\gamma}{b+\rho_1\gamma}\right) - \mathrm{E}_1\left(\frac{\gamma}{b+\rho_1\gamma}\,(b+1-\rho_1)\right)\right) = l$$

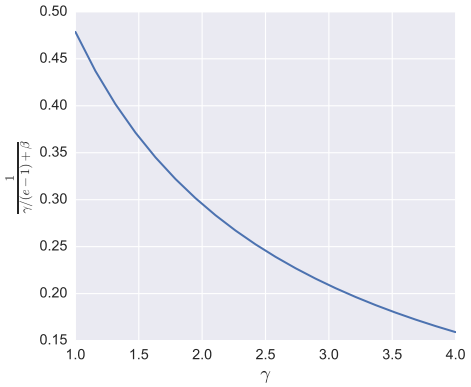
Figure 3.7A, depicts γ as a function of l for $b = 100$ and three different values of ρ_1.

Example 3. We consider $h(u) = 1 - \frac{1}{u+1}$, i.e., $H(U) = n - \mathbf{tr}((I+U)^{-1})$, relevant to the problem of A-optimal experiment design.[8] In this case $\nabla H(U) = (I+U)^{-2}$ which does *not* satisfy the PSD diminishing returns assumption, Assumption 2. As such, to obtain any

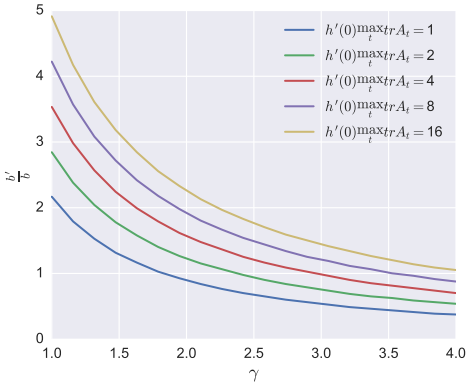
[8]This example also covers $n\epsilon^{-1} - \mathbf{tr}((\epsilon I + U)^{-1})$ with $\epsilon > 0$. This is because by replacing A_t with A_t/ϵ, one can equivalently write the problem with $H(U) = n - \mathbf{tr}((I+U)^{-1})$ as the objective.



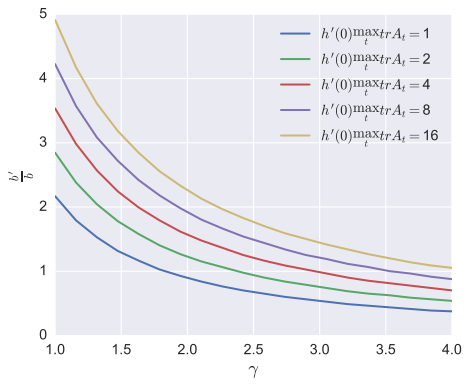
(A) competitive ratio bound vs γ



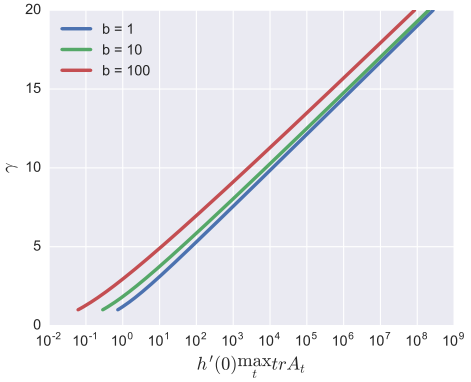
(B) competitive ratio bound vs γ



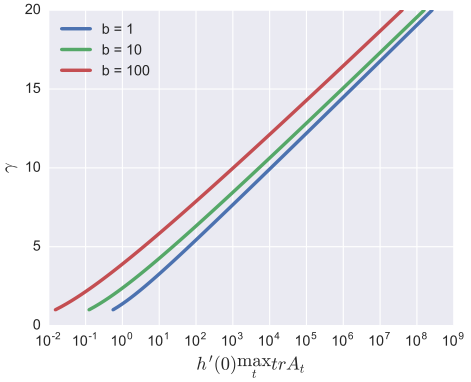
(C) b'/b vs. γ, with γ as a free parameter.



(D) b'/b vs. γ, with γ as a free parameter.



(E) γ given as a solution of (3.48) as a function of l



(F) γ given as a solution of (3.48) as a function of l

Figure 3.5: In (A),(C),(E), $h(u) = \log(u+1)$. In (B),(D),(F), $h(u) = 1 - 1/(u+1)$.

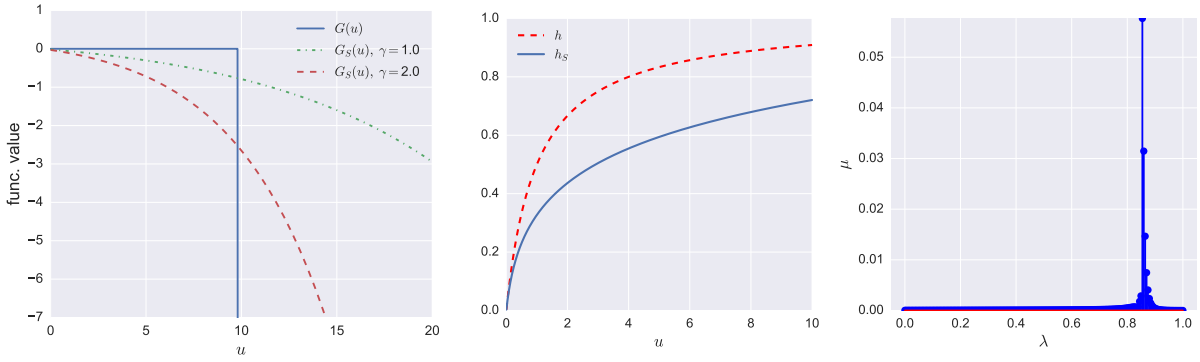


Figure 3.6: (a) Examples of G_S, and (b) h_S and the corresponding measure μ, when $h(u) = 1 - 1/(u+1)$.

competitive ratio bound using the approach presented, we must smooth H to some H_S that does satisfy Assumption 2. In this case $h^*(y) = 2\sqrt{y} - y + 1$. The optimal $G'_S = z$ can be expressed in terms of E_2 (defined in (3.57)) as

$$z(u) = -\frac{\gamma}{b}\exp\left(\frac{\gamma}{b}(u+1)\right)\left(\mathrm{E}_2\left(\frac{\gamma}{b}\right) - \frac{1}{b+1}\mathrm{E}_2\left(\frac{\gamma}{b}(u+1)\right)\right)$$

$$z(u) = \begin{cases} -\frac{\gamma}{b}\exp\left(\frac{\gamma}{b}(u+1)\right)\left(\mathrm{E}_2\left(\frac{\gamma}{b}\right) - \frac{1}{b+1}\mathrm{E}_2\left(\frac{\gamma}{b}(u+1)\right)\right), & u \in [0,b] \\ -l, & u > b \end{cases}$$

where γ is chosen so that $z(b) = -l$.

An example of h_S and G_S for this problem was given in Figure 3.6. Figure 3.5F depicts γ as a function of l for different values of b. Notice that γ scales logarithmically with l. In Figure 3.5A, we plot the competitive ratio bound of Theorem 9 vs γ. As mentioned if l is not available to the algorithm designer, γ can be treated as a free parameter that should be chosen by the designer. In that case the budget violation depend on γ and in Figure 3.5D, we plot b'/b versus γ (as a free parameter), when $h(u) = 1 - 1/(u+1)$..

For the sequential algorithm $G_S(u) = \int_0^u z(s)\,ds$ is given by:

$$z(u) = -\frac{\gamma}{(b+\rho_1\gamma)(e-1)}\exp\left(\frac{\gamma(u+1-\rho_1)}{b+\rho_1\gamma}\right)\left(\mathrm{E}_2\left(\frac{\gamma}{b+\rho_1\gamma}\right) - \frac{1}{b-\rho_1+1}\mathrm{E}_2\left(\frac{\gamma(u+1-\rho_1)}{b+\rho_1\gamma}\right)\right).$$

where γ satisfies:

$$\frac{\gamma}{(b+\rho_1\gamma)(e-1)}\exp\left(\frac{\gamma\,(b+1-\rho_1)}{b+\rho_1\gamma}\right)\left(\mathrm{E}_2\left(\frac{\gamma}{b+\rho_1\gamma}\right)-\frac{1}{b-\rho_1+1}\mathrm{E}_2\left(\frac{\gamma\,(b+1-\rho_1)}{b+\rho_1\gamma}\right)\right)=l$$

Figure 3.7B, depicts γ as a function of l for $b = 100$ and three different values of ρ_1.

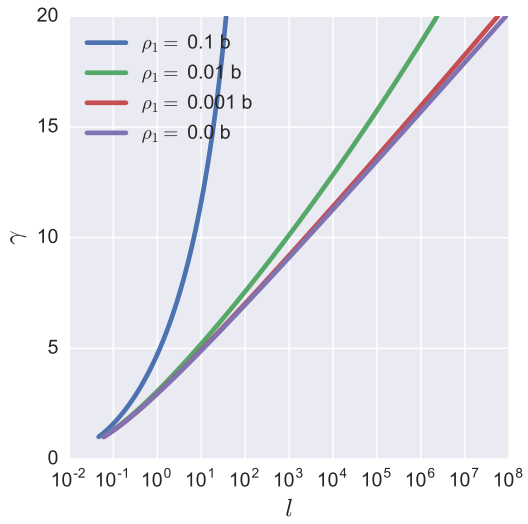
3.4 Related work in online optimization and learning

We discuss results and papers from two communities, computer science theory and machine learning, related to this work.

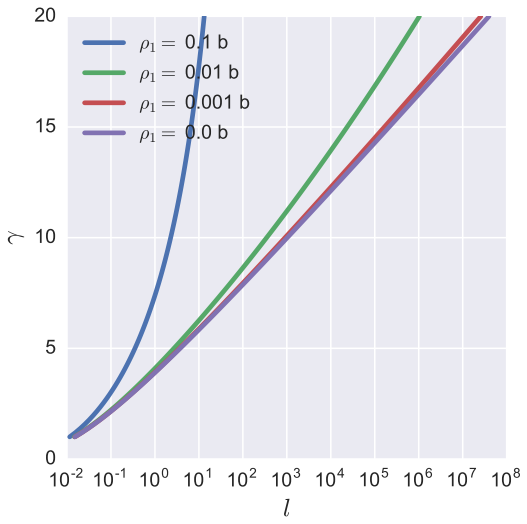
Online convex optimization. In [DJ12], the authors proposed an optimal algorithm for adwords with differentiable concave returns (see examples in section 3.2). Here, "optimal" means that they construct an instance of the problem for which competitive ratio bound cannot be improved, hence showing the bound is tight. The algorithm is stated and analyzed for a twice differentiable, separable $\psi(u)$. The assignment rule for primal variables in their proposed algorithm is explained as a continuous process. A closer look reveals that this algorithm falls in the framework of algorithm 3, with the only difference being that at each step, $(\tilde{x}_t, \tilde{y}_t)$ are chosen such that where $v_i : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is an increasing differentiable function given as a solution of a nonlinear differential equation that involves ψ_i and may not necessarily have a closed form. The competitive ratio is also given based on the differential equation. They prove that this gives the optimal competitive ratio for the instances where $\psi_1 = \psi_2 = \ldots = \psi_n$.

Note that this is equivalent of setting $\psi_{S_i}(u_i) = \psi(v_i(u_i)))$. Since v_i is nondecreasing ψ_{S_i} is a concave function. On the other hand, given a concave function $\psi_{S_i}(\mathbb{R}_+) \subset \psi_i(\mathbb{R}_+)$, we can set $v_i : \mathbb{R}_+ \mapsto \mathbb{R}_+$ as $v_i(u) = \inf\{z \mid \psi_i(z) \geq \psi_{S_i}(u)\}$. Our formulation in section 3.3.2 provides a *constructive* way of finding the optimal smoothing. It also applies to non-smooth ψ.

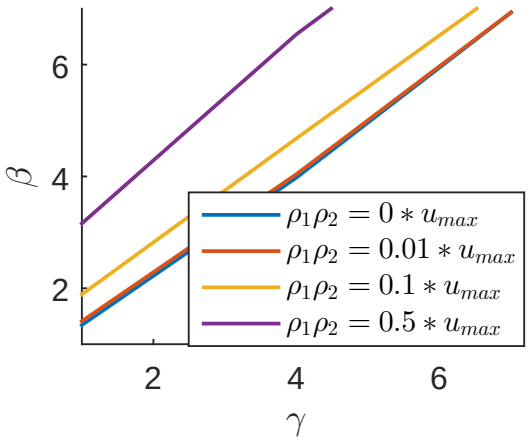
Recently, authors in [ACP14, BCG$^+$14, CHK15] have provided a primal-dual online algorithm for the dual problem (3.3) that corresponds to the non-monotone primal objective



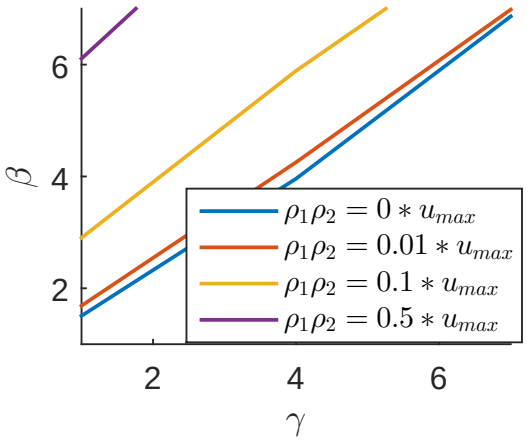
(A) γ vs. l for the sequential algorithm.



(B) γ vs. l for the sequential algorithm.



(C) β vs. γ for the sequential algorithm.



(D) β vs. γ for the sequential algorithm.

Figure 3.7: In (a) and (c), $h(u) = \log(u+1)$. In (b) and (d), $h(u) = 1 - \frac{1}{u+1}$.

$\psi\left(\sum_{t=1}^{m} A_t x_t\right) = \sum_{t=1}^{m} c_t^T x_t + G\left(\sum_{t=1}^{m} B_t x_t\right)$.The primal and dual updates in their algorithm are presented as a continuous update based on a differential equation. They assume that G^* is differentiable and that ∇G^* is monotone on $\mathbb{R}^n_+$, i.e., If $y \geq y'$, then $\nabla G^*(y) \leq \nabla G^*(y')$. In contrast, our assumption written in terms of G^* for a differentiable function will become: If $\nabla G^*(y) < \nabla G^*(y')$, then $y \geq y'$, which is not equivalent to the assumption in [BCG$^+$14]. When G is separable the two assumptions coincide and this algorithm is similar to algorithm 3 applied to the smooth function G_M whose conjugate is given by $G_M^*(y) = \frac{1}{\gamma}\sum_{i=1}^{n} \int_0^{y_i} G_i^{*\prime}(z) \log(1 - z/\theta)\, dz$. This smoothing coincides with Nesterov smoothing in the case of LP.

Online learning. As mentioned before, the dual update in Algorithm 6 is the same as in Follow-the-Regularized-Leader (FTRL) algorithm with $-\psi^*$ as the regularization. This primal dual perspective has been used in [SSS07b] for design and analysis of online learning algorithms. In the online learning literature, the goal is to derive a bound on *regret* that optimally depends on the horizon, m. The goal in the current chapter is to provide competitive ratio for the algorithm that depends on the function ψ. Regret provides a bound on the duality gap, and in order to get a competitive ratio the regularization function should be crafted based on ψ. A general choice of regularization which yields an optimal regret bound in terms of m is *not* enough for a competitive ratio argument, therefore existing results in online learning do not address our aim.

3.5 Relation with submodularity

In Section 3.2.3, we briefly mentioned the relationship between the diminishing return assumption over the non-negative orthant and submodularity. Here we elaborate on the relationship and draw connections to submodularity and submodular maximization.

3.5.1 An excursion into definitions.

Here, we review the definition of submodularity and its connection to the diminishing return property. Consider a lattice $(L, \vee, \wedge)$, where $\vee$ and $\wedge$ are the join and meet operation. In very general terms, a function f is submodular on a lattice $(L, \vee, \wedge)$ if

$$f(v \vee u) + f(v \wedge u) \leq f(v) + f(u). \tag{3.58}$$

for all u and v in L (see [Sch02, Chapter 49], [Top11] and [Sim14]). Perhaps the most important lattice with respect to which submodularity has been studied (and regularly exclusively defined on) is the lattice generated by the subsets of a ground set V. In this case the definition of submadularity requires:

$$f(T \cup S) + f(T \cap S) \leq f(T) + f(S).$$

for all $\{S, T\} \subset 2^V$ (reader interested in further reading on submodular set functions can see [Fuj05]). Every lattice determines a partially ordered set (*poset*), where the order $\leq$ is defined as:

$$u \leq v \text{ if and only if } u \vee v = v \text{ (equivalently, } u \wedge v = u). \tag{3.59}$$

Given this partial ordering, the diminishing return property for a function on a lattice can be defined as

$$f(x \vee d) - f(x) \geq f(y \vee d) - f(y) \text{ if } x \leq y \text{ and } d \wedge y \leq x, \tag{3.60}$$

where the partial ordering is given by (3.59). In the special case of the lattice $(2^V, \cup, \cap)$ this is equivalent to,

$$f(T \cup \{e\}) - f(T) \leq f(S \cup \{e\}) - f(S) \quad \text{if } S \subset T,\ e \notin T.$$

Proposition 4. *A function f satisfies the diminishing return assumption given in* (3.60) *with respect to the partial ordering induced by a distributive lattice if and only if it is submodular over the lattice.*

Proof. **Necessity.** Suppose $x \leq y$ and d is such that $d \wedge y \leq x$. Let $u = y$ and $v = x \vee d$. We have:

$$u \vee v = y \vee (x \vee d) = y,$$

by the fact that the lattice is distributive:

$$u \wedge v = y \wedge (x \vee d) = (x \wedge y) \vee (y \wedge d) = x.$$

This shows that (3.60) follows from (3.58).

Sufficiency. First note that $u \wedge v \leq u$. Now, inequality (3.58) can be derived from (3.60) by setting $d = v$, $x = u \wedge v$, and $y = u$. □

On the other hand the diminishing return property can be defined for a function f defined on a partially ordered vector space as:

$$f(x + d) - f(x) \geq f(y + d) - f(y) \text{ if } x \leq y \text{ and } d \geq 0. \tag{3.61}$$

This is specially of interest, since important examples like the space of symmetric matrices with semidefinite ordering is a partially ordered vector space but the partial ordering does not lend itself to a lattice structure.

If a partially ordered vector space has an order structure that is also a lattice (called a Reisz space) such as $\mathbf{R}^n$ with partial order $\leq_{\mathbf{R}^n_+}$, then similar to Proposition 3.7, one can relate this definition of DR to submodularity over the order-lattice of the Reisz space (see [Top11] for a more general version).

Proposition 5. *If a function f over a Reisz space satisfies the diminishing return assumption given in* (3.61)*, then it is submodular over the order-lattice.*

Proof. Inequality (3.58) can be derived from (3.61) by setting $x = u \wedge v$, $y = u$ and $d = v - u \wedge v$ and noting that:

$$y + d = u + v - u \wedge v = u \vee v \tag{3.62}$$

See [AB03, Theorem 1.3] for the proof of (3.62). □

It is important to note that the diminishing return assumption (3.61) in general is not necessary for (3.58). In order to find an example to show this one needs to look no further than $(\mathbf{R}^n, \leq_{\mathbf{R}^+_n})$. The function $f(x) = x_1^2 - x_1x_2$ satisfies (3.58) but does not satisfy (3.61).

In order to draw connection the DR Assumption 2, we define another notion of DR which we show is equivalent to (3.61). Given a function ψ on $(\mathbf{R}^n, \leq_K)$, recall that $\nabla\psi(u; d)$ is the directional derivative of ψ at u in the direction of d (denoted by $\psi(u; d)$). Now, we can state the following proposition:

Proposition 6. *Let ψ be is function on $\mathbf{R}^n$ with a given partial ordering $\leq$, then ψ satisfies* (3.61)*, if and only if*

$$\psi'(u; d) \leq \psi'(v; d) \text{ if } u \geq v \text{ and } d \geq 0 \tag{3.63}$$

for all $u, v, d \geq 0$

Proof. **Necessity**. If $u \geq v$ and $d \geq 0$, by (3.61), we have:

$$\frac{\psi(u + sd) - \psi(u)}{s} \leq \frac{\psi(v + sd) - \psi(v)}{s},$$

For all $s \geq 0$. Now by taking the right limit we get the desired result.

Sufficiency. If $u \geq v$ and $d \geq 0$, then we can write:

$$\psi(u + d) - \psi(u) = \int_0^1 d^T\psi'(u + sd, d)\, \mathbf{d}s \leq \int_0^1 d^T\psi'(v + sd; d)\, \mathbf{d}s = \psi(v + d) - \psi(v)$$

□

When ψ is a concave function Assumption 2 ensures (3.63) holds with partial ordering $\leq_K$. This follows from the fact that

$$\psi'(u; d) = \inf_{\xi \in \partial\psi(u)} \langle \xi, d \rangle$$

When ψ is differentiable, then Assumption 2 is equivalent to (3.63). In another word,

$$\psi'(u; d) \leq \psi'(v; d) \text{ for all } d \in K \Leftrightarrow \nabla\psi(u) \leq_{K^*} \nabla\psi(v).$$

The diminishing return condition on the non-negative orthant has been used in [BB17] as a sufficient condition for submodularity of a set function given by composition of a continuous function and a vector of polymatroid functions.

3.5.2 Fifty shades of greedy

The greedy algorithm has been well studied in the context of matroids and combinatorial optimization. The celebrated result by Nemhauser et al. [NWF78] proves that the greedy algorithm achieves $1 - 1/e$ approximation ratio for maximization of a non-decreasing submodular set function subject to cardinality constraint. The problem statement is as follows:,

$$\begin{aligned} \text{maximize} \quad & f(S) \\ \text{subject to} \quad & |S| \leq m, \quad S \subset V \end{aligned} \tag{3.64}$$

where f is a submodular set function and $|S|$ denotes the cardinally of set S. The greedy algorithm for submodular maximization is as follows:

Algorithm 6 Greedy for submodular maximization

Initialize $S_0 = \varnothing$

for $t \leftarrow 1$ to m **do**

$a_t \in \arg\max_{s \in V} f(S_{t-1} \cup \{a\})$

$S_t \leftarrow S_{t-1} \cup \{a_t\}$

end for=0

Let $S^* = \{a_1^*, \ldots, a_m^*\}$ be an optimal solution for problem 3.64. The proof for the approximation ratio is as follows:

$$f(S_{t-1} \cup \{a_t\}) - f(S_{t-1}) \geq \frac{1}{m} \sum_{t=1}^{m} f(S_{t-1} \cup \{a_t^*\}) - f(S_{t-1}) \quad \text{since } a_t \in \arg\max_{s \in V} f(S_{t-1} \cup \{a\})$$

On the other hand, using the diminishing return property, we have:

$$\begin{aligned} f(S_{t-1} \cup S^*) - f(S_{t-1}) &= \sum_{t=1}^{m} f(S_{t-1} \cup \{a_1^*, \ldots, a_t^*\}) - f(S_{t-1} \cup \{a_1^*, \ldots, a_{t-1}^*\}) \\ &\leq \sum_{t=1}^{m} f(S_{t-1} \cup \{a_t^*\}) - f(S_{t-1}) \end{aligned}$$

Combining the two inequalities, we get:

$$\begin{aligned} f(S_{t-1} \cup \{a_t\}) - f(S_{t-1}) &\geq \frac{1}{m} \left(f(S_{t-1} \cup S^*) - f(S_{t-1}) \right) \\ &\geq \frac{1}{m} \left(f(S^*) - f(S_{t-1}) \right) \qquad \text{by monotonicity of } f \end{aligned} \tag{3.65}$$

We can take a weighted sum of all the inequalities for $t = 1, \ldots, m$,

$$\sum_{t=1}^{m} w_t f(S_{t-1} \cup \{a_t\}) - w_t \left(1 - \frac{1}{m}\right) f(S_{t-1}) \geq \sum_{t=1}^{m} \frac{w_t}{m} f(S_{t-1} \cup S^*) \tag{3.66}$$

$$\geq \sum_{t=1}^{m} \frac{w_t}{m} f(S^*) \tag{3.67}$$

By setting $w_t = (1 - 1/m)^{1-t}$ and assuming $f(\varnothing) \geq 0$, the left hand side of (3.66) simplifies to:

$$\begin{aligned} \sum_{t=1}^{m} w_t f(S_{t-1} \cup \{a_t\}) - w_t \left(1 - \frac{1}{m}\right) f(S_{t-1}) &= \sum_{t=1}^{m} w_t f(S_t) - w_{t-1} f(S_{t-1}) \\ &= w_m f(S_m) - w_0 f(S_0) \\ &\leq \left(1 - \frac{1}{m}\right)^{1-m} f(S_m) \end{aligned} \tag{3.68}$$

while the right hand side of (3.66) simplifies to:

$$\sum_{t=1}^{m} \frac{w_t}{m} f(S^*) = \left(1 - \frac{1}{m}\right)^{1-m} - \left(1 - \frac{1}{m}\right) f(S^*) \tag{3.69}$$

Combining (3.68), (3.69), and (3.66), we get

$$f(S_t) \geq \left(1 - \left(1 - \frac{1}{m}\right)^{m}\right) f(S^*)$$

Several works in the submodular maximization literature have focused on extending the analysis of greedy approximation ratio to broader range of problems such as monotone submodular maximization subject to matroid constraint and knapsack constraints (for example, see [CCPV07, DS06, Von08, Svi04]). In [Von08], the author applies the greedy algorithm to the multilinear extention of a submodular function accompanied with pipage rounding

[AS04], to achieve $1 - 1/e$ approximation ratio for submodular optimization subject to matroid constraints. The multilinear extention of a submodular set function satisfies the diminishing return assumption (3.63).

The greedy algorithm, by the virtue of making sequential decisions without backtracking, is a natural candidate for online optimization. Perhaps the closest link between online optimization and submodular maximization comes through the competitive ratio analysis of the online greedy algorithm under the i.i.d. setting (that is, when $(A_1, F_1), \ldots, (A_m, F_m)$ are i.i.d sampled from an unknown distribution). Competitive ratio analysis of the online greedy algorithm under the i.i.d. setting is very similar to the $1 - 1/e$ approximation analysis we reviewed above.

Devanur et al. [DJSW11] first proved that the greedy algorithm achieves $1-1/e$ competitive ratio under the i.i.d. model. The adwords problem can also be viewed as submodular welfare maximization problem [GS07], which in turn is an special case of submodular maximization with a matroid constraint. Kapralov et al. [KPV] extended the analysis of the online greedy algorithm under the i.i.d. setting to the online submodular welfare problem. Here, we give the proof of $1 - 1/e$ competitive ratio of greedy algorithm under the i.i.d. setting for Algorithm 3 applied to problem (3.1) in order to point to the similarities with the previous proof of the approximation ratio of the greedy algorithm.

Consider the simultaneous algorithm (Algorithm 3). Assume that $A_1, \ldots, A_m$ are uniformly sampled with replacement from a set $\{B_1, \ldots, B_m\}$ and to simplify the derivations, we assume that all $F_t = F$ for some convex compact set. We give the competitive ratio for concave ψ that satisfies Assumption 2 and is monotone. We let $z_1^*, \ldots, z_m^*$, be the optimal solution to the following problem:

$$\begin{aligned} \text{maximize} \quad & \psi\left(\sum_{t=1}^{m} B_t x_t\right) \\ \text{subject to} \quad & x_t \in F, \quad \forall t \in [m], \end{aligned} \tag{3.70}$$

First, note that we have:

$$\psi\left(\sum_{t=1}^{m} B_t z_t^*\right) \geq E\left[\max_{x_t \in F} \psi\left(\sum_{t=1}^{m} A_t x_t\right)\right]. \tag{3.71}$$

This follows from the fact for any $(x_1, \ldots, x_m) \in F^m$, by Jensen's inequality:

$$\begin{aligned} E\left[\psi(\sum_{t=1}^{m} A_t x_t)\right] \leq \psi\left(\sum_{t=1}^{m} E[A_t] x_t\right) &= \psi\left(\sum_{t=1}^{m}\left(\frac{1}{m}\sum_{t=1}^{m} B_t\right) x_t\right) \\ &= \psi\left(\sum_{t=1}^{m} B_t\left(\frac{1}{m}\sum_{t=1}^{m} x_t\right)\right) \end{aligned} \tag{3.72}$$

We use $\mathcal{A}_t$ to denote the sigma algebra generated by $A_1, \ldots, A_t$. We have:

$$E\left[\psi\left(\sum_{s=1}^{t} A_s \tilde{x}_s\right) - \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s\right)\middle| \mathcal{A}_{t-1}\right] \geq E\left[\psi\left(\sum_{s=1}^{t} A_s \tilde{x}_s\right)\middle| \mathcal{A}_{t-1}\right] - \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s\right)$$

By the primal assignment rule

$$= \frac{1}{m}\sum_{t=1}^{m} \max_{x \in F_t} \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s + B_t x\right) - \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s\right)$$

$$\geq \frac{1}{m}\psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s + B_t z_t^*\right) - \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s\right)$$

By the diminishing return assumption

$$\geq \frac{1}{m}\left(\psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s + \sum_{t=1}^{m} B_t z_t^*\right) - \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s\right)\right)$$

By montonicity of ψ

$$\geq \frac{1}{m}\left(\psi\left(\sum_{t=1}^{m} B_t z_t^*\right) - \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s\right)\right) \tag{3.73}$$

The last inequality is the counterpart to (3.65). The rest of the proof follows from similar steps to the previous proof and culminates in:

$$\begin{aligned} E\left[\psi\left(\sum_{s=1}^{t} A_s \tilde{x}_s\right)\right] &\geq \left(1 - \left(1 - \frac{1}{m}\right)^m\right) \psi\left(\sum_{t=1}^{m} B_t z_t^*\right) \\ &\geq \left(1 - \left(1 - \frac{1}{m}\right)^m\right) E\left[\max_{x_t \in F} \psi\left(\sum_{t=1}^{m} A_t x_t\right)\right] \qquad \text{By (3.72).} \end{aligned}$$

To compare and contrast this with the worst-case analysis, we write a primal only analysis for the worst-case competitive ratio of the simultaneous algorithm. Recall that $(x_1^*, x_2^*, \ldots, x_m^*)$ is the offline optimal solution for (3.1) when $(A_1, F_1), \ldots, (A_m, F_m)$ are generated by the adversary. By the primal assignment rule, We can write:

$$\psi\left(\sum_{s=1}^{t} A_s\tilde{x}_s\right) - \psi\left(\sum_{s=1}^{t-1} A_s\tilde{x}_s\right) \geq \psi\left(\sum_{s=1}^{t-1} A_s\tilde{x}_s + A_t x_t^*\right) - \psi\left(\sum_{s=1}^{t-1} A_s\tilde{x}_s\right)$$

Here, as opposed to the previous proof, the lower bound on the function value improvement at time t only depend on x_t^*. To relate it to the optimal value we first take the sum of the inequalities for all t, and then use the diminishing return assumption:

$$\begin{aligned}
\psi\left(\sum_{s=1}^{m} A_s\tilde{x}_s\right) &\geq \sum_{t=1}^{m}\left(\psi\left(\sum_{s=1}^{t-1} A_s\tilde{x}_s + A_t x_t^*\right) - \psi\left(\sum_{s=1}^{t-1} A_s\tilde{x}_s\right)\right) \\
&\quad \text{By the diminishing return property} \\
&\geq \sum_{t=1}^{m}\left(\psi\left(\sum_{s=1}^{m} A_s\tilde{x}_s + A_t x_t^*\right) - \psi\left(\sum_{s=1}^{m} A_s\tilde{x}_s\right)\right) \\
&\quad \text{By the diminishing return property} \\
&\geq \psi\left(\sum_{s=1}^{m} A_s\tilde{x}_s + \sum_{t=1}^{m} A_t x_t^*\right) - \psi\left(\sum_{s=1}^{m} A_s\tilde{x}_s\right) \\
&\quad \text{By monotonicity of } \psi \\
&\geq \psi\left(\sum_{t=1}^{m} A_t x_t^*\right) - \psi\left(\sum_{s=1}^{m} A_s\tilde{x}_s\right).
\end{aligned}$$

This yields

$$\psi\left(\sum_{s=1}^{m} A_s\tilde{x}_s\right) \geq \frac{1}{2}\psi\left(\sum_{t=1}^{m} A_t x_t^*\right)$$

As we discussed in Section 3.2.1, for a non-deacrsing ψ, we have $\bar{\alpha}_\psi \geq -1$ and the competitive ratio bound given by our primal-dual analysis is $1/(1-\bar{\alpha}_\psi) \geq \frac{1}{2}$.

In order to improve upon the $1 - 1/e$ approximation ratio of the greedy algorithm for submodular maximization subject to a cardinality constraint for special functions, one can

use the notion of total curvature [CC84] defined as:

$$c_f = 1 - \min_{S \subset V, j \in S/V} \frac{f(S \cup \{j\}) - f(S)}{f(\{j\})}.$$

In [CC84], It has been shown that the approximation ratio of the greedy algorithm is lower bounded by $\frac{1}{c_f}(1 - \exp(-c_f))$. This result has been also extended to general matroid constraint in [Von10].

To get the same result for competitive ratio under i.i.d. setting, one can analogously define the curvature for a continuous function ψ as:

$$c_\psi = 1 - \min_{\{u,v\} \subset K} \frac{\psi(u+v) - \psi(u)}{\psi(v)}$$

Modifying the primal proof given above, one gets a competitive ratio bound of $\frac{1}{c_\psi}(1 - \exp(-c_\psi))$ for the i.i.d. setting, by the and $1/(1 + c_\psi)$ for the worst-case setting.

In the i.i.d. setting the, the inequality (3.73) is modified to:

$$E\left[\psi\left(\sum_{s=1}^{t} A_s \tilde{x}_s\right) - \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s\right) \middle| \mathcal{A}_{t-1}\right] \geq \frac{1}{m}\left(\psi\left(\sum_{t=1}^{m} B_t z_t^*\right) - c_\psi \psi\left(\sum_{s=1}^{t-1} A_s \tilde{x}_s\right)\right) \tag{3.74}$$

The weight w_t in the weighted sum of the inequalities is modified to $w_t = \left(1 - \frac{c_\psi}{m}\right)^{1-t}$.

If the function ψ plateaus ($\nabla\psi(u) = 0$ for some u), then $c_\psi = -\alpha_\psi = 1$ while for a linear function $c_\psi = -\bar{\alpha}_\psi = 0$. In general, for a concave function ψ, we have

$$\begin{aligned}\bar{\alpha}_\psi = \min_{u \in K} \frac{\langle \nabla\psi(u), u\rangle}{\psi(u)} - 1 &\geq \min_{u \in K} \frac{\langle \nabla\psi(u), u\rangle}{\langle \nabla\psi(0), u\rangle} - 1 \\ &\geq \min_{\{u,v\} \subset K} \frac{\langle \nabla\psi(u), v\rangle}{\langle \nabla\psi(0), v\rangle} - 1 = -c_\psi\end{aligned}$$

There are functions for which $c_\psi > -\alpha_\psi$. Consider $\psi(u) = \sqrt{u}$. Then we have $\bar{\alpha}_\psi = -0.5$ (worst-case competitive ratio bound of 2/3), while $c_\psi = 1$ (worst-case competitive ratio bound of 1/2). Even if one restricts the domain of u in the definition of c_ψ in order to ms

Chapter 4

CLUSTERING AND CATEGORIZATION OF V4 CELLS

4.1 Introduction

The ventral pathway of the primate visual system is involved in form processing and object recognition, but only the early stages of this pathway are well understood. While most neurons in the primary visual cortex are tuned for orientation and spatial frequency, cells in area V4, an intermediate stage along this pathway, appear to be selective for more complex visual patterns and have been more difficult to characterize. Cells in area V1 are well known to be selective for the local orientation and spatial frequency in small patches of an image; however, cells in area V4 have larger receptive fields, appear to be selective for more complex visual patterns and have been more difficult to characterize [KT94, GBVE93, PC01]. V4 neurons have been shown to respond to aspects of shape [PC01], shading [AJG09], texture [AJG08], and color [Zek83], and it is likely that additional important axes of representation remain to be discovered. Several quantitative models have been proposed in the literature to explain shape selectivity in V4 neurons. This includes the angular-position and curvature (APC) model [PC01], the hierarchical-max (H-Max) network model [CKP$^+$07], and the spectral receptive field (SRF) model [DHG06].

In many studies attempting to understand V4, the stimulus design and analysis were driven by a specific underlying hypothesis and model. To avoid inherent pitfalls in this approach, and discover novel encoding dimensions, we have taken a more model-free approach based on clustering of the cells to reanalyze responses of V4 neurons to a set of simple shapes. We accumulated data from 3 studies (6 animals) conducted by Pasupathy and Connor [PC01], Popovkina et al. [PBP], Oleskiw et al. [ONP16] for a total of 272 V4 neurons. These studies all used the same stimulus set (proposed in [PC01]) consisting of 362 stimuli (51 shapes at up to 8 rotations). We defined a rotation-invariant metric between pairs of cells based on their responses to the shape set such that cells with shape preferences that are rotated version of each other will be separated by a very small distance. Using multiple clustering methods, we found 4 major clusters that included $\sim$25% of the cells. Independently, We determined how well the cells within each cluster were fit by the published models

for V4 - the APC, SRF, and H_Max models - and examined how well the clusters could be distinguished in the parameter space of these models. This latter step is important, because it points to new aspects of the selectivity that the models did not capture, and determines the clustering simply recapitulated diversity within the parameter space of existing models. Overall, our clustering identified novel categories of units that were not distinguished based on the parameters of the existing models tested so far, and we were able to identify clear intuitive dimensions, e.g., tuning for stimulus area and isoperimetric quotient, that did distinguish these clusters. The existence of V4 cells that are tuned to shape area has not been discovered before in the literature. These results demonstrate that the approach proposed here can reveal basic insights about shape representation that were previously overlooked, and they strongly support our hypothesis that there are clusters of neurons with qualitatively different visual selectivities that call for neurons to be appropriately classified before they can be finely characterized.

4.2 Materials and Methods

Neuronal data set and Experimental Methods. We combined extraceullar single-unit data from three studies of V4 neurons in awake, fixating macaque monkeys (six animals overall). This includes 109 cells from [PC01], 43 cells from [PBP] and 63 cells from [ONP16]. See the original studies for details of surgical procedures and behavioral training. All of these studies used the stimulus set proposed by Pasupathy and Conner [PC01], which we will refer to as the PC2001 shapes. Figure 4.1 shows the full set of stimuli, where each of 51 shapes are presented at 8 rotations (or fewer, for shapes that have rational symmetries). This stimulus set is designed to capture the behavior of boundary-curvature tuned cells in the V4 area and has been subsequently used in multiple studies on V4 neurons [KESFP14, BP12].

Visual stimuli were presented for 500 ms and were presented typically 5 per fixation with blank periods of typically 250 ms in between. The neuronal response to a stimulus was defined to be the mean firing rate across repeats of that stimulus. Firing rate was calculated in a window of length 500 ms where the stimulus was presented. Typically, there were 5-10

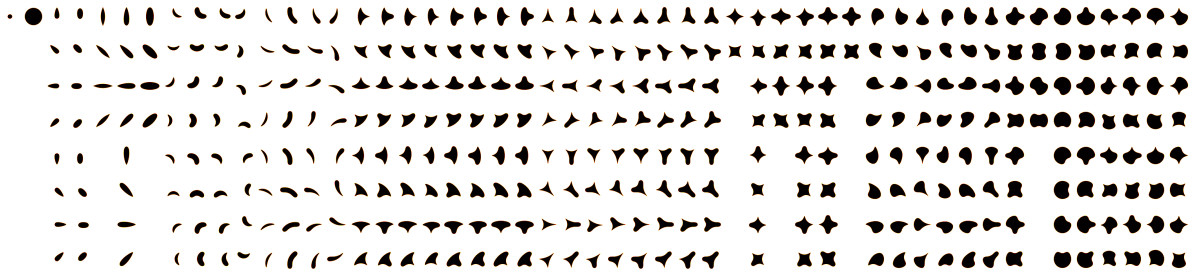


Figure 4.1: The PC2001 shape set. There are 52 distinct shapes (across top row) that were shown at up to 8 rotations (down columns) in the three studies from which data was re-analyzed here. For example, the small and large circle (left two shapes) have only 1 unique rotation. Shapes are drawn to scale.

repeats of each stimulus, but sometimes as few as 3.

Cluster analyses. To cluster V4 neurons on the basis of their responses to the shape set, we first define a rotation-invariant disimilarity measure between any pair of cells based on their responses such that cells with shape preferences that are rotated version of each other will be separated by a very small distance. Let $G = \{g^0, g^1, \ldots, g^7\}$ be the set of 8 rotations in the plane such that g^k maps a shape to $k \times \pi/4$ rotated version of it. Given vectors x and y each comprised of the average firing rate of a cell in response to all the 362 stimuli, we define the dissimilarity measure between x and y as:

$$\mathrm{d}(x, y) = \frac{1}{2}(1 - \max_{g \in G} \mathrm{corr}(x_g, y)) \tag{4.1}$$

Here x_g denotes a permuted version of vector x in which the response to a shape is replaced by the response to the shape rotated by g.

We used a technique known as Density Based Spatial Clustering of Applications with Noise (DBSCAN) [EKS+96], which is suitable for clustering in the presence of outliers. This method assigns one of three possible labels to each point in the dataset: core, reachable and outlier. Each cluster consists of at least one core point and any reachable points, i.e., those within a critical distance of a core point in that cluster. Outliers are points that belong to

no cluster. DBSCAN takes two parameters that control the number of clusters that will be found: ϵ and $MinPts$. Each point in a cluster has distance less than or equal to ϵ to at least one core point in that cluster. The parameter $MinPts$ determines the minimum number of points needed to form a cluster.

Rotation Variance Index. We defined a rotation variance index to quantify how much the responses of a cell depended on the rotation of shapes in the stimulus set. For this, we only consider responses to 44 shapes in Figure 1 that are shown at eight rotations, i.e., that do not have any rotational symmetry. We first form a 44 by 8 matrix M that contains the response of a cell to the 44 shapes in 8 rotations. Then the rotation variance index is defined as the square of the ratio of the nuclear norm to the Frobenius norm of that matrix:

$$\frac{\left(\sum_{i=1}^{8} \sigma_i(M)\right)^2}{\sum_{i=1}^{8} \sigma_i(M)^2}, \tag{4.2}$$

where σ_i is the i^{th} singular value. For a non-zero response matrix M, this index takes values between 1 and 8, and a smaller index signals a higher degree of rotation invariance.

The angular position and curvature (APC) model. This model was proposed in [PC01], to capture the tuning of V4 cells for the curvature of the boundary of simple closed shapes. This model postulates that a cell is tuned for a specific boundary curvature at a specific angular position with respect to a coordinate system at the center of the shape. Curvature varies from concave (-1) to convex ($+2$), and angular position varies from 0-2π around the center of the shape (for details, see [PC01]). Mathematically, finding the parameters of this model requires fitting a Gaussian kernel in the two dimensional space of angular position and curvature. The APC has four parameters: angular position mean, angular position SD, curvature mean, and curvature SD. To fit this model to the response of a cell to the set of 362 shapes, each shape is first represented as 8 points in the two dimensional space of angular position and boundary curvature. The 8 points correspond to samples of boundary curvature of the shape at 8 different angular position's $\{0, \pi/4, \pi/2, \ldots, 7\pi/8\}$. The

response of the model to the a shape represented as $\{(x_1, y_1), \ldots, (x_8, y_8)\}$ in the angular position and curvature space is:

$$\max_i \exp\left(-\frac{(x_i - \mu_a)^2}{2s_a^2} - \frac{(y_i - \mu_c)^2}{2s_c^2}\right), \tag{4.3}$$

where $\{\mu_a, s_a, \mu_c, s_c\}$ are the angular position mean, angular position SD, curvature mean, and curvature SD. In other words, if we think of a shape as being defined by 8 major curvature features around its boundary, then the response of the model to that shape is determined by the feature that falls at the highest point on the 2D Gaussian tuning function (see [PC01], page 2508). We found these parameters for each cell by maximizing the correlation between the response vector of the cell and that of the model. We used this correlation (r-value) as the goodness-of-fit metric.

To capture the response of a cell that is tuned for different adjoining boundary curvatures (for example, a concavity adjoined by a convexity), a more complicated version of this model can be considered by representing the shapes in a higher dimensional space. In addition to the simple APC model, we consider a 4 dimensional APC model (4D APC). To calculate the response of this model to a shape, we first represent the shape with 8 point in a 4 dimensional space: $\{(x_1, y_1, v_1, w_1), \ldots, (x_8, y_8, v_8, w_8)\}$, where y_i, v_i and w_i are the boundary curvature at angular positions x_i, $x_i + \pi/4$ and $x_i - \pi/4$, respectively. The response of the model to the shape is given by

$$\max_i \exp\left(-\frac{(x_i - \mu_a)^2}{2s_a^2} - \frac{(y_i - \mu_{c_1})^2}{2s_{c_1}^2} - \frac{(w_i - \mu_{c_2})^2}{2s_{c_2}^2} - \frac{(v_i - \mu_{c_3})^2}{2s_{c_3}^2}\right). \tag{4.4}$$

SRF model. The Spectral Receptive Field (SRF) model is proposed in [DHG06] to model shape selectivity in V4. In this model the response of a cell to an image is a linear function of the power spectrum of the image. We fit this model based on the response of the cells to the set of 362 shapes. In order to prevent overfiting we use a LASSO [Tib96] problem

$$\text{minimize}_x \quad \|Ax - b\|_2^2 + \lambda \|x\|_1. \tag{4.5}$$

Where the regularization parameter is chosen by cross-validation.

LN model. In this model the predicted firing rate is given by a nonlinearity applied to the output of a linear filter:

$$r(s) = f(k.s), \tag{4.6}$$

where f is a nonlinear function and k denotes a linear filter. Under the assumption that spikes are generated according to an inhomogeneous Poisson process, the log-likelihood function is given by:

$$LL(k) = \sum_{s \in S} R(s) \log(r(s)) - r(s), \tag{4.7}$$

where $R(s)$ is the observed firing rate in response to the shape stimulus s in S, the set of all shape stimuli. To estimate k, we solve the following optimization problem:

$$\text{maximize}_k \quad LL(k) - \lambda \|\mathcal{L}(k)\|_1, \tag{4.8}$$

where $\mathcal{L}$ is the discrete two-dimensional Laplacian operator and λ is the regularization parameter.

4.3 Results

In this section we first describe the clusters that were found by the clustering analysis. We then compare the clusters based on their fits to different models, and characterize the tuning of the cells in the largest clusters.

4.3.1 Clustering

We carried out a cluster analysis of V4 data from several animals based on the rotation invariant dissimilarity measure (see methods). We found that several clusters tended to appear regardless of the details of the clustering methods. Here we discuss the results of clustering by DBSCAN algorithm. We varied the parameter ϵ and in Figures 4.2A we show the number of cells in each cluster as a function ϵ when $MinPts = 4$. Figures 4.2B shows the same plot when $MinPts = 5$. The same color labels have been used in the two plots for the

	Total	Red	Brown	Blue	Green	Purple	Gray
Pasupathy & Conner [PC01]	109	9	5	9	4	3	4
Oleskiw et al. [ONP16]	63	6	1	14	0	0	0
Popovkina et al. [PBP]	43	1	1	8	0	3	0
Kosai et al. [KESFP14]	100	7	1	0	0	1	1
Total	315	23	8	31	4	7	5

Table 4.1: Summary of Clustering

clusters that share the majority of their members. When $MinPts = 4$, this algorithm finds 5 clusters which we assign a color name to: Red, Blue, Purple, Green, and Brown. When $MinPts = 5$ the algorithm does not find the purple cluster. The total number of clustered cells in the five clusters mentioned above is 65, which amounts to %30 of the total number of cells. As evident in this figure, the blue cluster is the first cluster that forms meaning that the initial members are closer together than the initial members of clusters that form later. The red cluster is the second cluster that is found by the algorithm. In addition to the 215 cells that were used in this clustering, we had access to 100 cells from [KESFP14] that used only a subset of size 40 of the PC2001 shapes. We trained a multi-class SVM classifier based on the initial clustering of the original 215 cells to classify these 100 cells into the 6 clusters. Table 4.1 summarizes the result of the clustering and the classification.

To further visualize the clustering, we have depicted the dissimilarity matrix for the clustering with $MinPts = 4$ in Figure 4.2C. The (i, j) entry of the matrix is dissimilarity measure and can take values between 0 and 1. The five clusters form the five blocks on the diagonal that appear deeper blue, indicating a tendency to have low dissimilarity values. Figure 4.2D, shows the similarity matrix only for the clustered cells. When cells cluster together, the most direct interpretation is to say that they share a substantial degree of their tuning preferences, once rotation of the visual stimulus is factored out. As is apparent in

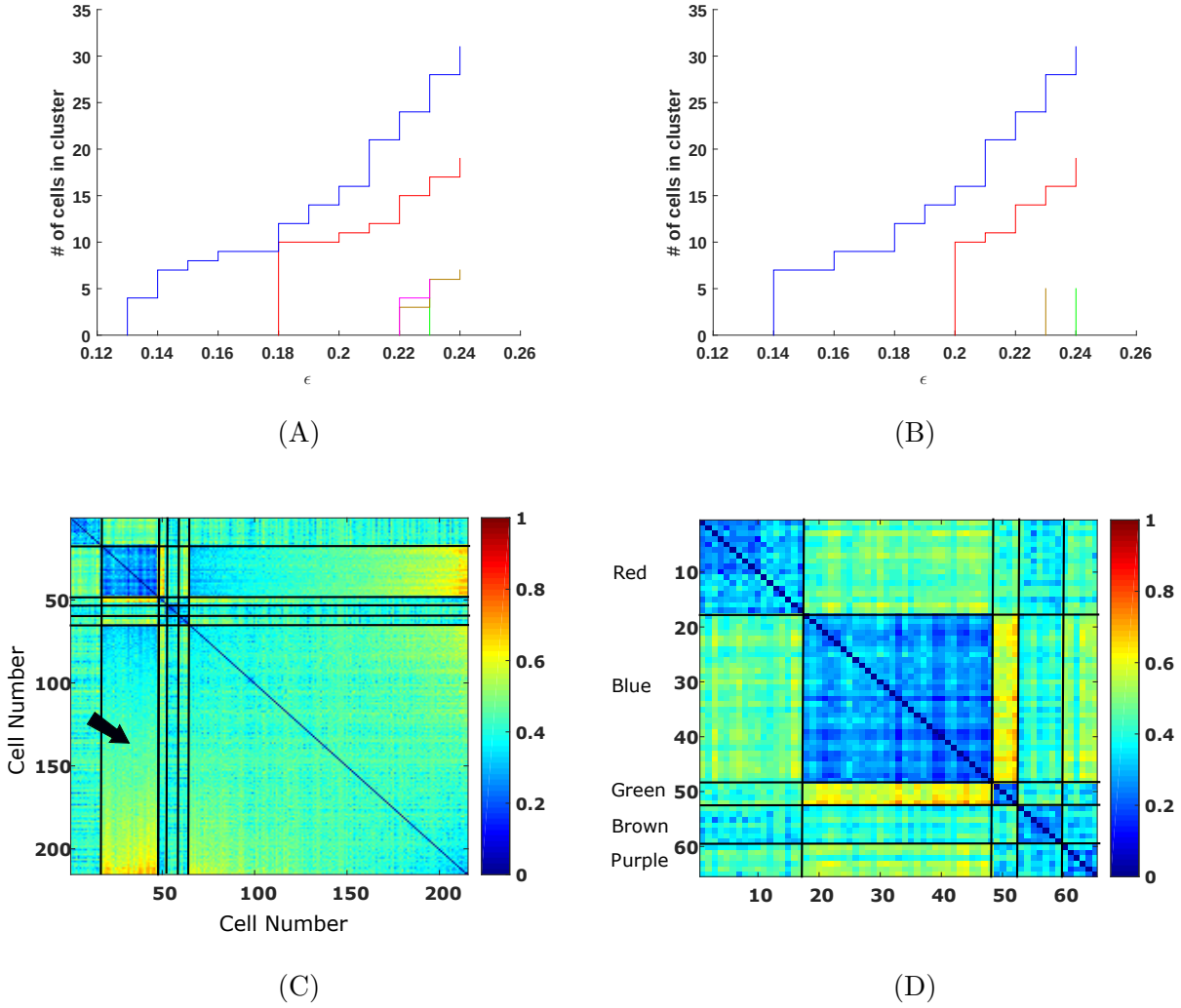


Figure 4.2: Visualization of the result of clustering with DBSCAN. **A** and **B**, The number of points in each cluster vs. the value of ϵ for $MinPts = 4$ (**A**) and $MinPts = 5$ (**B**). **C**, The similarity matrix for all the cells. **D**, The similarity matrix for all the clustered cells.

Figure 4.2D, blocks that adjoin clusters often show low similarity (yellow to red), indicating separation between clusters. In particular, the Red cluster vs. Blue cluster, and Blue cluster vs. Green cluster, the latter being particularly strong. Interestingly, many of the unclustered cells show dissimilarity to the Blue cluster. After observing this trend, we sorted the unclustered cells based on their dissimilarity to Blue cluster. The arrow in the figure shows the rectangle associated with the relationship between Blue and the unclustered cells, and it is apparent that it contains more yellow-red color than other regions in the matrix.

4.4 Red Cluster and The APC Model

Having identified several clusters of cells based on their raw responses, we wanted to determine whether these clusters were associated with interpretable and distinct visual selectivity. One obvious place to begin is to compare the selectivity of neurons in each cluster to the features that the stimulus set was originally designed to explore - the boundary curvature of the shapes. To do this, we examined the parameters and goodness of the fits to the APC model. Figure 4.3A shows the Pearson's r-value between the predicted response by the APC model and the observed response of the cell for all the clusters and outliers. We found that the distributions of APC r-values varied substantially across clusters. In particular, the Red cluster had the best fit to the APC model on average, and it was significantly better than the blue, green and unclustered units (t-test, $p < .005$ in all cases). This suggests that clusters vary with respect to their tuning for boundary curvature. Figure 4.3B shows the SD vs. mean for the angular position parameter, whereas Figure 4.3C shows SD vs. mean for the curvature parameter. Two clear distinctions emerge here. First, the Blue cluster has many cells with angular position SD far larger than those in any other cluster (Fig. 4.3B). This distinction is particularly obvious between Red and Blue clusters because they have the most elements. Second, the Red cluster has a large fraction of members for which the curvature mean (Fig 4.3C) lies at or above the value of 1 (sharp convexities), whereas no other clusters do. In this way, Red is particularly distinct from Blue, which has the lowest curvature means of all clusters. Overall, Blue has wide SDs for angular position and is relatively poorly fit

by the APC model (Fig 4.3A), suggesting that it has many members that do not embody the idea of angular-position tuning. Red, on the other hand, has consistently narrow tuning for angular position, often prefers sharp convexities, and has the highest r-values on average. In addition, it turns out that almost all of the example cells shown in papers on the APC model have used Red-group cells. Thus, Red group is an exemplary bunch of APC-tuned units, whereas Blue group may be more interested in some other other property, as we will see below.

We observe that the red cluster comprises of two subgroups with respect to the APC parameters: a subgroup with negative curvature mean (in $[-.33, 0]$) and a subgroup with large positive curvature mean (in $[.8, 1.5]$). To investigate the these two subgroups, we consider one example cell in each subgroup of the red cluster, cell #109 and cell #174. The APC model parameters for these two cells are given in Table 4.2 and the APC tuning functions are plotted in Figure 4.4A. The dissimilarity measure between the two cells is .22 (r-value of .56) which is achieved by the permutation of the response vector of cell #174 corresponding to 270° rotation of the response of cell #174 (See Materials and Methods). The dissimilarity measure between the predicted responses by the APC models for the two cells is .28 (r-value of .44) which is also achieved by the same permutation of the response vector of cell #174. Figure 4.4 shows the response of these two cells to the shape set with response vector of cell #174 permuted according to the 90° rotation of the shapes. As it is evident by the response plots, the response of the two cell are similar for the shapes for which a sharp point (large positive curvature) is adjoined by a mild concavity (negative curvature with small magnitude). Now since the curvature mean for cell #109 is 1.38 and curvature mean for cell #174 is -0.07, this artifact in the shape set creates correlation between the responses of two cells once rotation is factored out. This can be further corroborated by calculating the partial correlation between the responses of these two cells with the responses of the corresponding APC models removed. Let x and y be the response vectors corresponding to

cell	angular position mean	angular position SD	curvature mean	curvature SD
174	45°	17°	1.38	0.41
109	0°	28°	−0.07	0.1

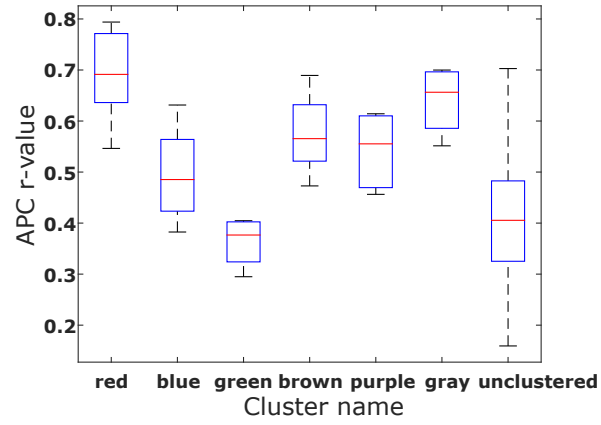
Table 4.2: APC parameters for two cells in the red cluster

cell #109 and to cell #174. Then we have

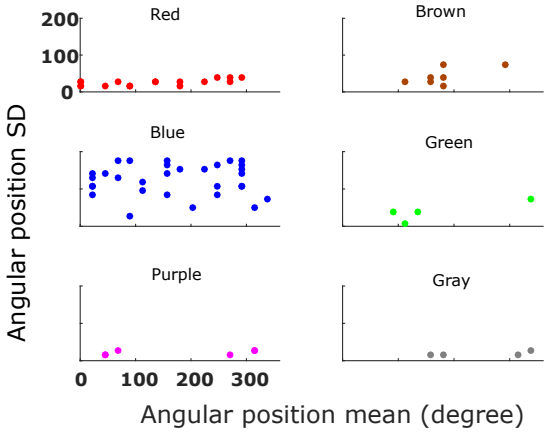
$$\mathrm{corr}(x - \bar{x}, y_{g^2} - \bar{y}_{g^2}) = -0.057, \tag{4.9}$$

where $\bar{x}$ and $\bar{y}$ are the predicted response by the APC models for the two cells.

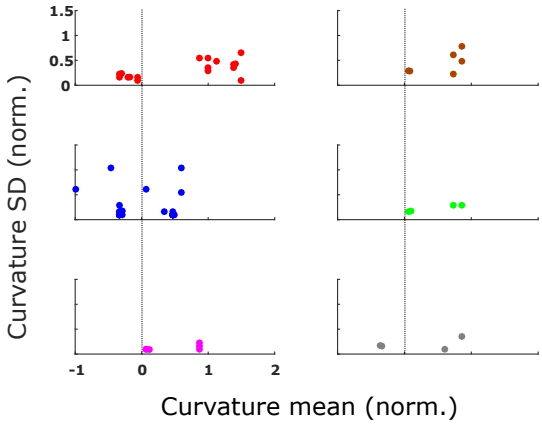
To understand to what degree our clusters might simply be different parts of the APC space, we generated responses from a set of models that covered the space and classified them based on our neuronal clusters. The criteria used for classification of the APC models is the same as the one used in DBSCAN. An APC model belongs to a cluster if the dissimilarity measure between the model and one of the core points in that cluster is less than ϵ. The APC model has four parameters. However, since our dissimilarity measure factors out rotation, we have fixed the angular position mean for all the APC models to 0 degree and considered 7000 models over a range of angular position SD, curvature mean, and curvature SD. Among 7000 APC models analyzed, %60 were clustered into one of the clusters of Red, Blue, Purple, Green, and Brown (Figure 4.5A). The majority of clustered models were labeled as Red cells (Figure 4.5B), and most of the models that were labeled as blue have large angular position SD. The models that are labeled as red also divide into two subgroups. We fit each model in the positive curvature subgroup with the models in the negative curvature subgroup and vice versa. On average the r-value is .55, this corresponds to a dissimilarity measure of .225.



(A)

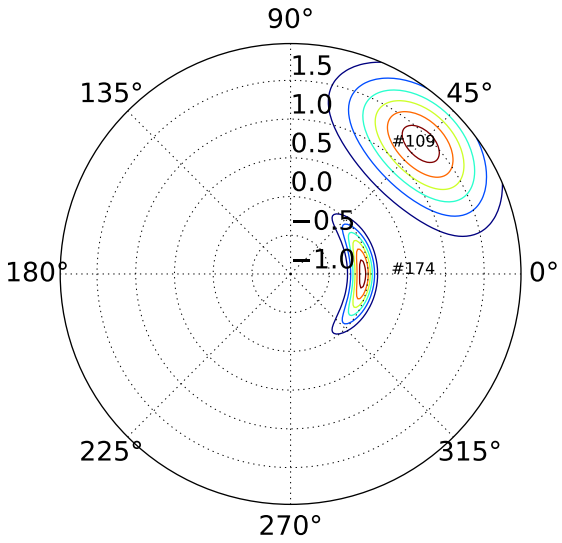


(B)



(C)

Figure 4.3: The APC parameters for the cells. **B**, The angular-position mean parameter and the angular-position variance parameter of the APC model for the cells in the six clusters.



(A)

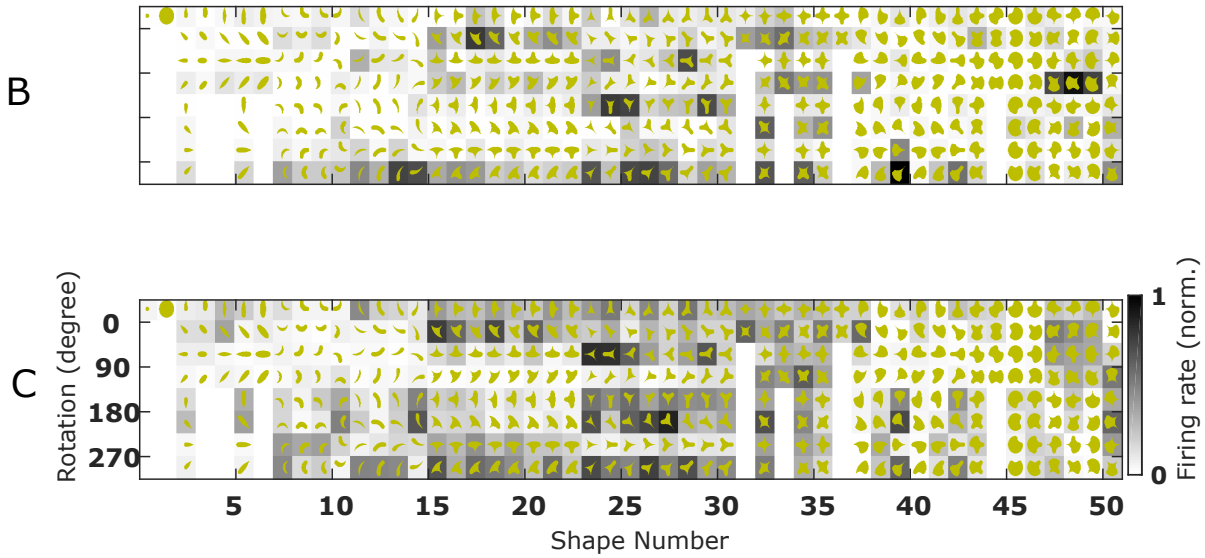
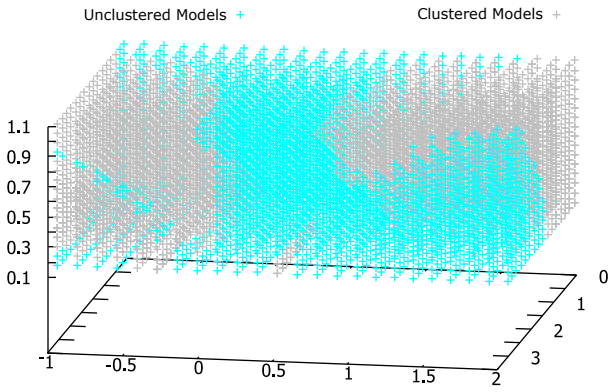
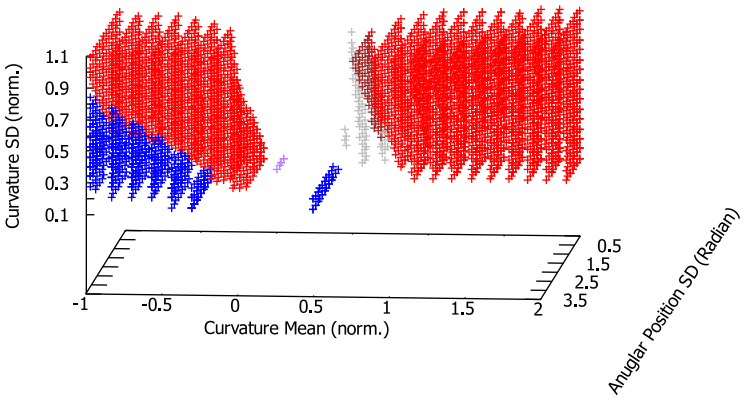


Figure 4.4: The response of cell #109 (**B**) and cell #174 (**C**) to the shape set. The response vector of cell #174 has been permuted in manner corresponding to 270° rotation in the shape set to maximize correlation with the response vector of cell #109. **A**, The contours of the APC tuning functions for cell #174 and #109. Angular position is represented by the θ axis and curvature is represented by the r axis.



(A)



(B)

Figure 4.5: **B**, APC models classified into the 5 clusters. **A**, Classified and unclassified APC models.

4.5 Blue Cluster and The SRF Model

Another model that has been applied to V4 is the SRF model [DHG06, OPB14], which uses a weighted sum across the Fourier spectrum of the visual stimulus to predict the neuronal response. Figure 4.6A compares SRF model r-value and the APC model r-value for all the clustered cells. The distinction between the nature of blue cluster and the red cluster becomes more evident here since the red cells are almost exclusively better fit by the APC model, while all the blue cells are better fit by the SRF model. In fact, among all the cells (clustered and outliers) the cells with the highest SRF model r-value are in the blue cluster. The green cells, which were the worst fit by the APC model 4.3A, are substantially better-fit by the SRF model and appear localized in this parameter space as well.

4.5.1 Tuning for shape area.

In addition to models motivated by past studies of V4, we also examined the clusters in terms of some fundamental features of the visual stimuli. One that ended up being particularly informative was shape area. Figure 4.7A shows, separately for each cluster, the average response to each shape vs. the area of the shape. The blue cluster appears to be driven by shape area. Since the area of a shape is given by the magnitude of the Fourier coefficient at frequency zero it should not be surprising that the SRF model fits these cells well. Figure 4.7A, shows the response to the 362 shapes by the area of the shapes. We fit a linear model to the response vector of each cell in the blue cluster with shape area as the independent variable. Figure 4.8A plots the r-value for the area model against that for the APC model for all the cells in the blue cluster. The y-coordinate of each circle represents the r-value between the predicted response vector by the linear model and the observed responses of each cell while the x-coordinate represents the r-value for the APC fit. All but three cells fall above the line of equality, confirming that tuning for shape area better explains the response of these cells in comparison to the APC model. Figure **??** shows the response of an example cell (cell #165) in the blue cluster. We have plotted the residual vector (actual

response - predicted response) of the linear model and the APC model for this example cell. The linear models have small residuals, while the APC models have relatively large residuals especially on a set of shapes that have low positive boundary curvature and small area. The curvature mean parameter for this cell is 0.47 which explains the strong predicted response of the APC model to these shapes. To further investigate the hypothesis on the area tuning of blue cluster we have computed the Rotation Invariance Index for all the clusters Figure 4.8B. Note that a higher number shows less rotation invariance.

To test for other plausible explanations of the tuning of blue cluster, we considered other parameters that are correlated to shape area over this set of shapes. The first quantity is the percentage of boundary having low positive curvature (normalized curvature in [0,.2]). This quantity takes larger values for big circle and many larger stimuli that have large curves. Figure 4.7B, shows the response vs. this quantity for all the six clusters. Note that some of the boomerang-shaped stimuli might provide the key to telling area apart from low-positive curvature. The second parameter is the isoperimetric quotient which is a measure of roundedness of the shape and is defined as

$$\text{Isoperimetric quotient} = \frac{4A}{L^2}$$

Where A is the area and L is the boundary length of the shape. Figure 4.7C shows the average response vs. isoperimetric quotient for all the six clusters. Note that this quantity takes values between 0 and 1 with the maximum achieved by circle. The large and small circles are the the key to telling area apart from isoperimetric quotient since both have the same isoperimetric quotient and yet the blue cells respond strongly to the large circle and very weakly to the small circle.

In summary, the blue cluster appears to be related to a trend whereby the V4 neurons tended to respond better to shape area than to the angular position of some boundary feature. This is consistent with the SRF model fits being better, because the SRF can account for area, as area is the spectral value for zero frequency.

4.6 The Remaining Clusters

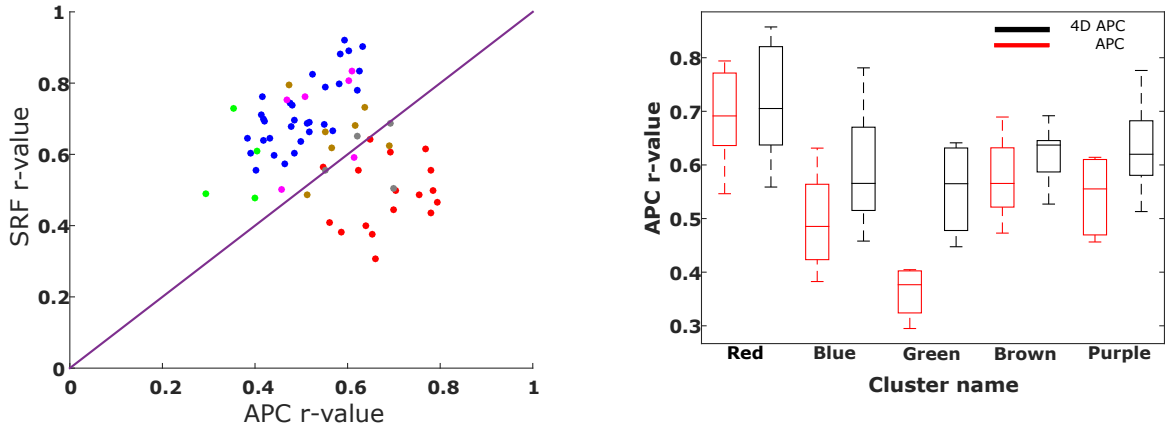
In this section we discuss our observations on the remaining clusters, i.e., Brown, Purple, Green and Gray.

The purple cluster. This cluster comprises of 6 cells. In Figure 4.13, we have depicted the response of three cells in the purple cluster (cells #83, #92, and #210). As it is evident by this figure, these cells respond strongly to a few shapes namely shape #3-#7. While the APC model for cell #83 predicts a strong response to shape #3-#7, it also predicts a strong response to shapes such as shape # 12 - # 23, but the the cell does not respond strongly to these shapes.

The brown cluster. The Brown cluster includes 7 cells. Figures 4.15A and 4.15B show the linear filter in the LN model for two example cells in the brown cluster. For comparison, in Figures 4.15C and 4.15D, we have depicted the linear filter for the two example red cells discussed in section 4.4. These linear filters confirm that these neurons possess a strong inhibitory region at one corner of the receptive field next to an elongated excitatory region that goes thorough the center of the receptive field.

The green cluster. Recall that the green cluster shows a strong negative correlation with the blue cluster (Figure 4.2D). In fact, the green cluster shows a decline in response relative with shape size (Figure 4.7A). In figure 4.6B, we have compared the APC r-value and 4D APC r-value for the clusters. The 4D APC models shows the largest improvement relative to the APC model for the green cluster.

The gray cluster The gray cluster includes 4 cells. After the red cluster it is the best fit cluster by the APC model (Figure 4.3A). The APC parameters for cells in the gray cluster fall in the same range of as those for the cells in the red cluster (Figures 4.3B and4.3C). However, there are differences between the response of cells in the red cluster and the gray



(A) (B)

Figure 4.6: H_max and SRF models. **A**, the comparison between the SRF model r-value and the APC model r-value for all the clustered cells.

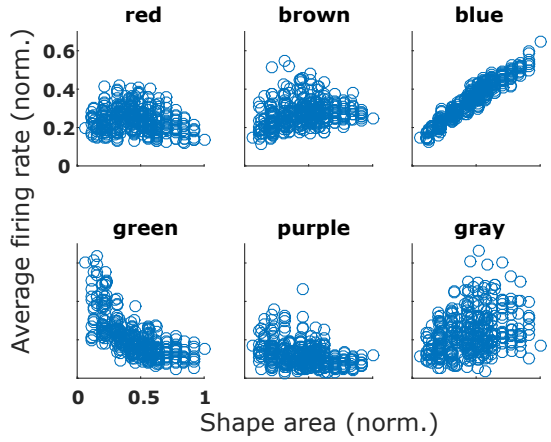
cluster. For example, the response of the cells in the gray cluster is mildly correlated by the shape area while for red cluster there is no such correlation (Figure 4.7A).

4.7 Adaptive Stimulus Sampling For Cell Classification

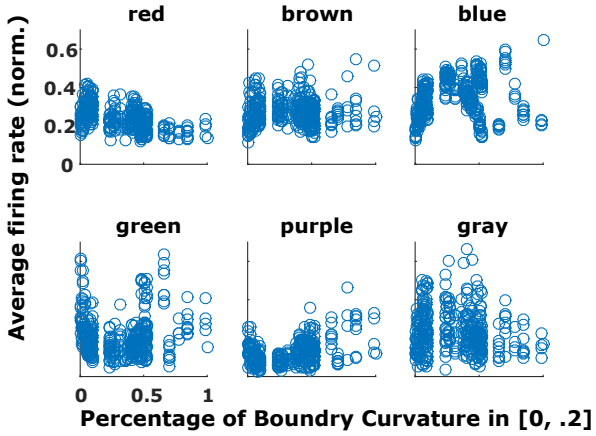
In this section, we consider adaptive greedy algorithms for selecting experiments in order to rapidly classify a cell in one of the clusters that we found. The algorithm that we consider is greedy algorithm and in nature similar to the algorithms considered in Section 3. In fact, greedy D-optimal experiment design, discussed in section 3.4 can be derived as a special case of the algorithm we consider here.

We first set up the notation and introduce the algorithm in generality and then specialize it to our problem.

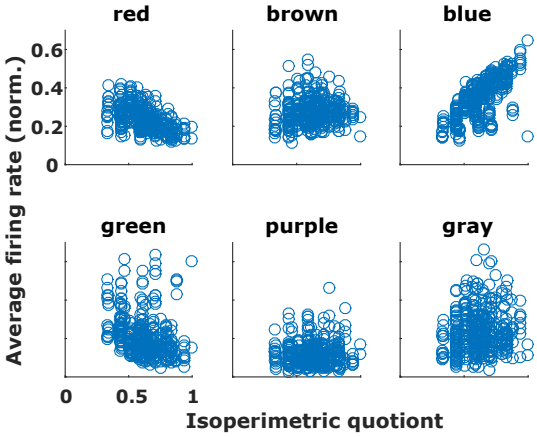
Let $\mathcal{H}$ be a set of hypotheses. We use S to denote the set of stimulus. The response of a cell to a stimulus s, is denoted by the random variable r_s. For the observed response we use bold font: $\mathbf{r}_s$. For any $T \subset S$, we define $r_T = \{r_t \mid t \in T)$ and $\mathbf{r}_T = \{\mathbf{r}_t \mid t \in T\}$.



(A)

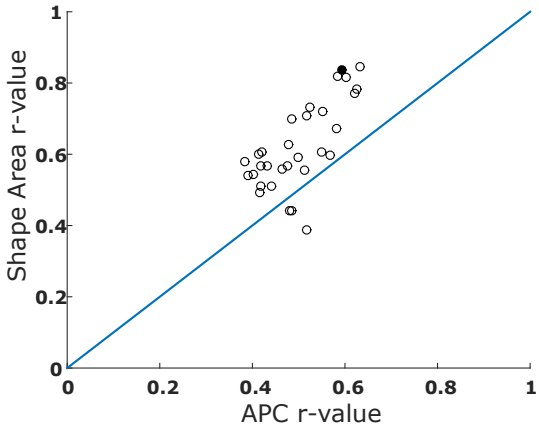


(B)

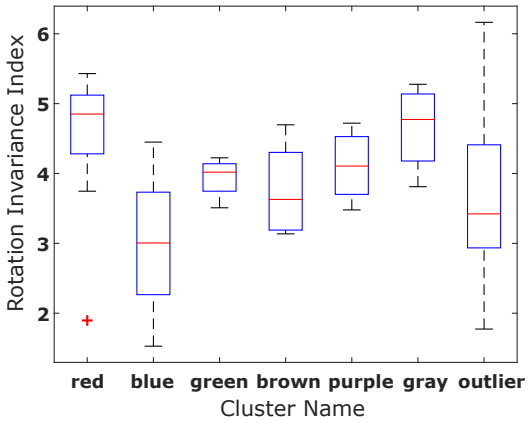


(C)

Figure 4.7: **A**, The average response of the cells in each cluster to a shape vs. the area of the shape. The y-coordinate of each circle represents the average response of the cells in the

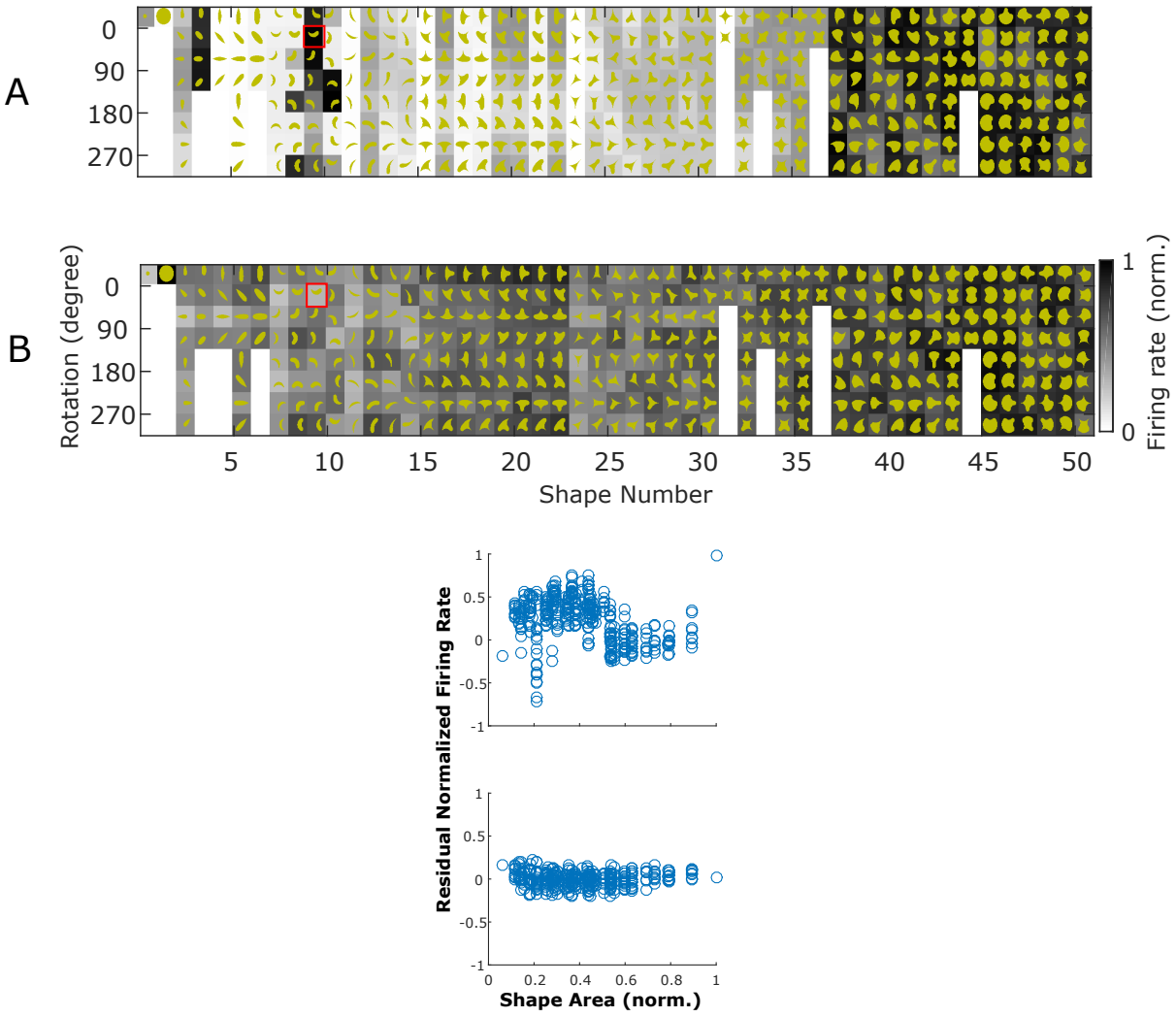


(A)



(B)

Figure 4.8: **A**, The r-value between the response vector of a cell to shape set and shape area vs. the APC model r-value for all the cells in the blue cluster. The filled point is from the example cell of Figure 4.9. **B**, Residual plot for the APC model (first row) and for the linear model as a function of shape area (second row) for 5 core cells in the blue cluster. **C**, The rotation invariance index for all the clusters and outliers.



(A)

Figure 4.9: **A**,The response of an example blue cell #165 to the shape set. B, the predicted responses of APC model cell #165. The APC parameters are: AP mean = 22°, AP SD = 107°, CR mean = 0.47, CR SD = 0.1. **C**, Residual plot for the APC model (top) and for the linear model as a function of shape area (bottom) for the same example blue cell.

In adaptive stimulus sampling after choosing stimuli in T, the next stimulus is chosen greedily in order to maximize a utility function. We use the notation $\psi(\mathbf{r}_T, t)$ to denote the utility function. This function quantifies the utility of using stimulus t after showing stimuli in T and observing the firing rate. A template for greedy stimulus selection is as follows:

Algorithm 7 Sequential Update

Set $T = \varnothing$.

for $i \leftarrow 1$ m **do**

$t \in \text{argmax}_{s \in S} \psi(\mathbf{r}_T, t)$

$T \leftarrow T \cup \{t\}$

This algorithm at each step picks a stimuli that maximizes the utility function. This algorithm is implement by the following class:

4.7.1 Utility functions

Mutual Information Utility

For any $T \subset S$ and $t \in S - T$, define

$$\psi(\mathbf{r}_T, t) = H(h \mid \mathbf{r}_T) - H(h \mid \mathbf{r}_T, r_t)$$

where H is the Shanon entropy. The quantity $\psi(T, t)$ captures the reduction in entropy when stimulus t is selected after observing the response to all the stimuli in T. The conditional entropy $H(h \mid \mathbf{r}_T, r_t)$ can be written as:

$$H(h \mid \mathbf{r}_T, r_t) = -\int dP(h, r_t \mid \mathbf{r}_T) \log(P(h \mid \mathbf{r}_T, r_t))$$

Now using Bayes rule we can compute the probability distributions involved in the definition of $H(h \mid \mathbf{r}_T, r_t)$:

$$\begin{aligned} P(h, r_t \mid \mathbf{r}_T) &= P(r_t \mid h, \mathbf{r}_T) P(h \mid \mathbf{r}_T) \\ &= P(r_t \mid h) P(h \mid \mathbf{r}_T) \qquad \text{Assuming that } r_t \text{ and } r_T \text{ are independent given } h \end{aligned}$$

For any $h \in \mathcal{H}$ we assume $P(r_t \mid h)$ is given as problem data while $P(h \mid \mathbf{r}_T)$ is updated recursively after each experiment. For example, after observing $\mathbf{r}_t$, we can calculate the conditional probability function $p(h|\mathbf{r}_{T\cup\{t\}})$ recursively according to the Bayes rule:

$$P(h|\mathbf{r}_{T\cup\{t\}}) = \frac{P(\mathbf{r}_t|h,\mathbf{r}_T)P(h|\mathbf{r}_T)}{\sum_{h\in\mathcal{H}} P(\mathbf{r}_t|h,\mathbf{r}_T)P(h|\mathbf{r}_T)} \tag{4.10}$$

The algorithm in more details:

Algorithm 8 Sequential Update

$P(r_s \mid h)$ for all $h \in \mathcal{H}$ and $s \in S$, $\quad \bar{P}$: the prior probability on $\mathcal{H}$

Set $T = \varnothing$;

for $j \leftarrow 1$ m **do**

Pick t that maximizes$\sum_{h\in\mathcal{H}} \bar{P}(h) \int dP(r_t \mid h)\ \log\left(\frac{P(r_t \mid h)\bar{P}(h)}{\sum_{h\in\mathcal{H}} P(r_t \mid h)\bar{P}(h)}\right)$

$T \leftarrow T \cup \{t\}$

Observe $\mathbf{r}_t$

For all $h \in \mathcal{H}$, update $\bar{P}$:

$\bar{P}(h) \leftarrow \frac{P(\mathbf{r}_t \mid h)\bar{P}(h)}{\sum_{h\in\mathcal{H}} P(\mathbf{r}_t \mid h)\bar{P}(h)}$

Maximum uncertainty

This utility function measures the uncertainty of the unobserved response to a stimulus given the observed responses to the previous stimuli. One popular measure of uncertainty in the literature is the entropy function: For any $T \subset S$ and $t \in S - T$, define

$$\psi(\mathbf{r}_T, t) = H(r_t|\mathbf{r}_T)$$

$$H(r_t \mid \mathbf{r}_T) = -\int dP(r_t \mid \mathbf{r}_T) \log(P(r_t \mid \mathbf{r}_T)),$$

where $P(r_t \mid \mathbf{r}_T)$ can be calculated as:

$$\begin{aligned} P(r_t \mid \mathbf{r}_T) &= \int P(r_t \mid h, \mathbf{r}_T)\ dP(h \mid \mathbf{r}_T) \\ &= \int P(r_t \mid h)\ dP(h \mid \mathbf{r}_T) \qquad \text{Assuming that } r_t \text{ and } r_T \text{ are independent given } h \end{aligned}$$

It has been shown in [SW97], that under the assumption that r_t is a deterministic function of h corrupted by additive noise this maximizing this utility function is equivalent to maximizing the information gain. However, this assumption does not hold in our setting. For example a simple Poisson firing rate model violates the additive noise assumption.

Adaptive stimulus sampling for cell classification

In this problem the set of hypothesis coincides with the set of V4 neuron classes. These classes are determined based on the clustering analysis. For each cluster h, we define an average response vector $x^{(h)}$ which is given by taking the average of the response vectors for all the cells in that cluster. Now, the corresponding conditional probabilities are defined as follows:

$$P(r_t \mid h) = \frac{e^{-x_t^{(h)}} \left(x_t^{(h)}\right)^{r_t}}{r_t!}$$

and

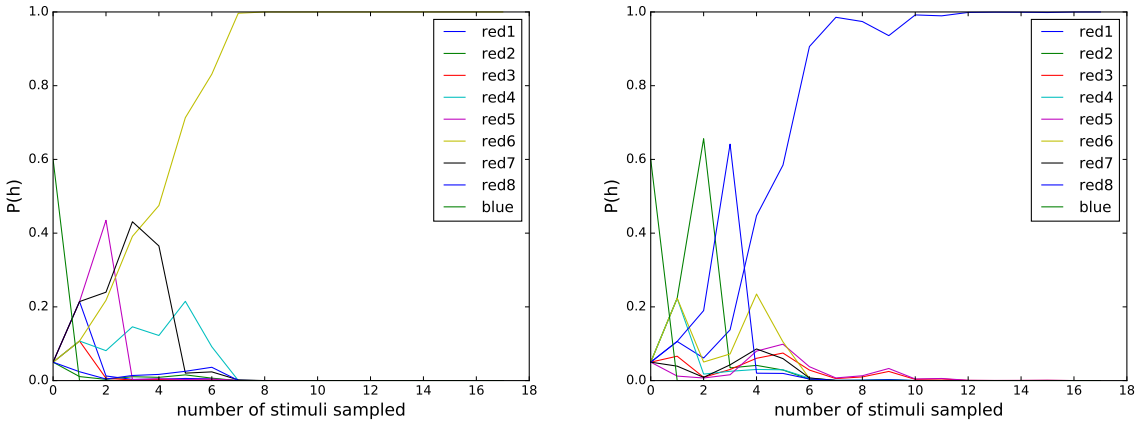
$$\bar{P}(h) = \frac{n_h}{\sum_{h' \in \mathcal{H}} n_{h'}},$$

where n_h is the number of cells in cluster h.

In our numerical examples, we consider 9 classes (Blue cluster and the red cluster under 8 rotations). We have simulated a cell in the red cluster with Poisson noise. Figure 4.10 shows examples of how the probability distribution on the set of hypotheses evolves during the course of the algorithm for the mutual information and uncertainty utility functions. In Figure 4.11, we compare the accuracy of classification for the two utility functions as a function of number of stimuli sampled by the algorithm.

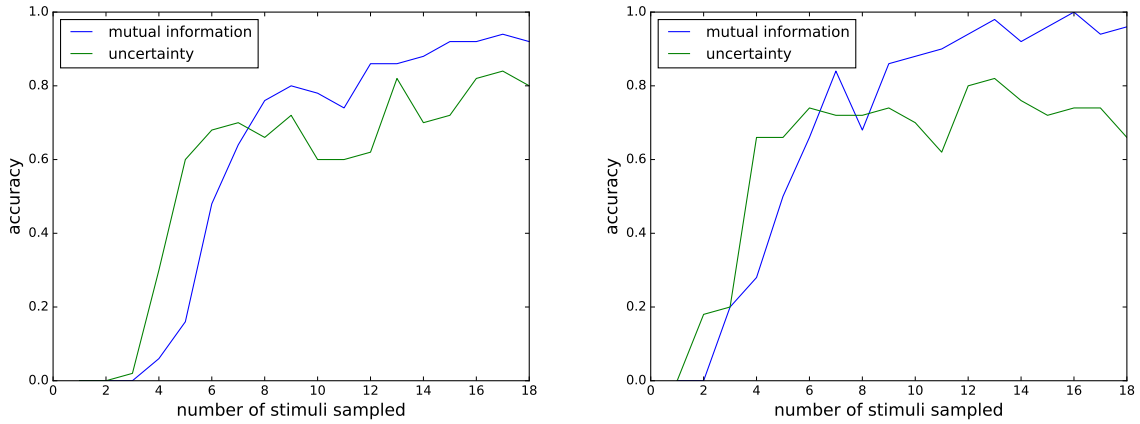
4.8 Discussion

As it was discussed in the previous section the red group comprises of cells for which the APC model provides an accurate predicted responses to the shape set. Figures 4.12 shows the histogram of the APC r-values for the clustered and outliers cells. From 32 cells with



(A) (B)

Figure 4.10: An example of performance of the algorithm for neuron classification problem.



(A) Ratio of instances when the right class has prob $> .9$ to total

(B) Ratio of instances when the right class has the highest probability among all

Figure 4.11: Comparison between the information and uncertainty utility functions.

APC r-value greater than .6, 9 cells are not in any cluster. The fact that not all APC cells are clustered would follow if there is a range of complexity of contour features encoded by APC cells. In particular, if some cells were selective for a simple feature, such as a sharp convexity at a tightly specified angle, with little regard for other features around the boundary, then all such cells would appear to like the same stimuli. However, if other cells had more elaborate requirements, say a conjunction of features at three positions, then the chance of finding two or more cells that had that preference could be far less. For example, say there were 4 levels of curvature feature: sharp and medium by convex and concave, and say there were 2 angular amounts to separate the three features, then there would be 4 x 2 x 4 x 2 x 4 ways to arrange these features, thus 256 different tuning functions, and we would not expect to find many matches across a few hundred neurons. Therefore, we speculate that the Red group contains APC cells with rather simple, elemental feature preferences with tuning that covers all rotations, and thus there are many cells in this group.

We also observed that the red cluster can be divided into two subgroup based on the sign of the curvature mean parameter in the APC model. This calls into design of a new stimuli set that can distinguish these two subgroups.

The blue group appears to be tuned for the size (area) of the stimulus. One important control would be to independent vary the size and luminance of the stimuli, to be sure that it is size, and not overall light.

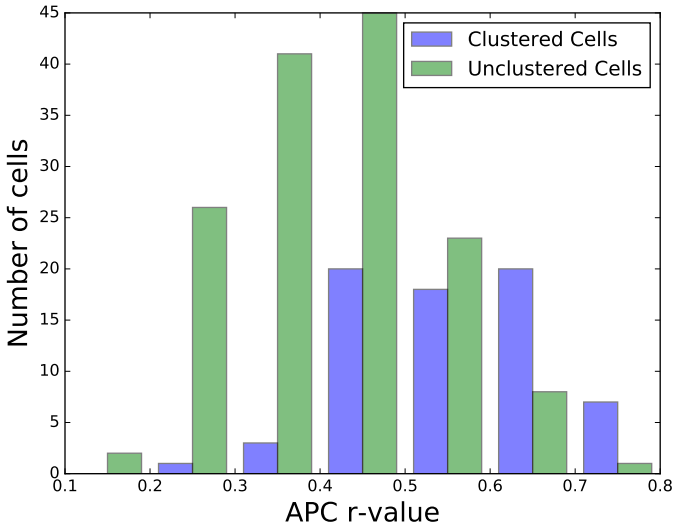


Figure 4.12: The histogram of APC r-values for clustered and unclustered cells.

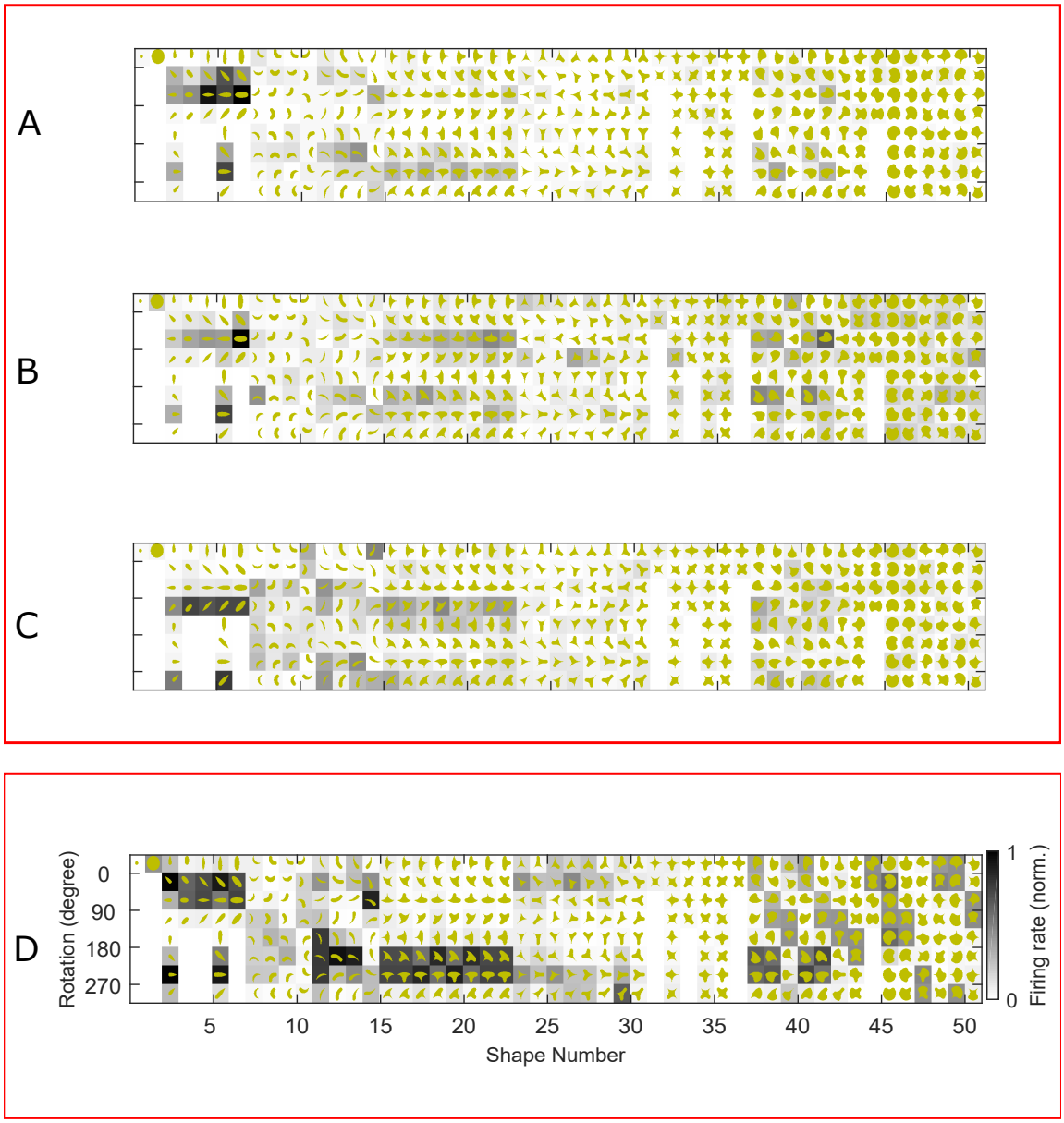


Figure 4.13: The response of three cells in the purple group, cell #83, cell #92 and cell #210, to the shape set (**A, B, C**). **D** The response of the APC model for cell #83.

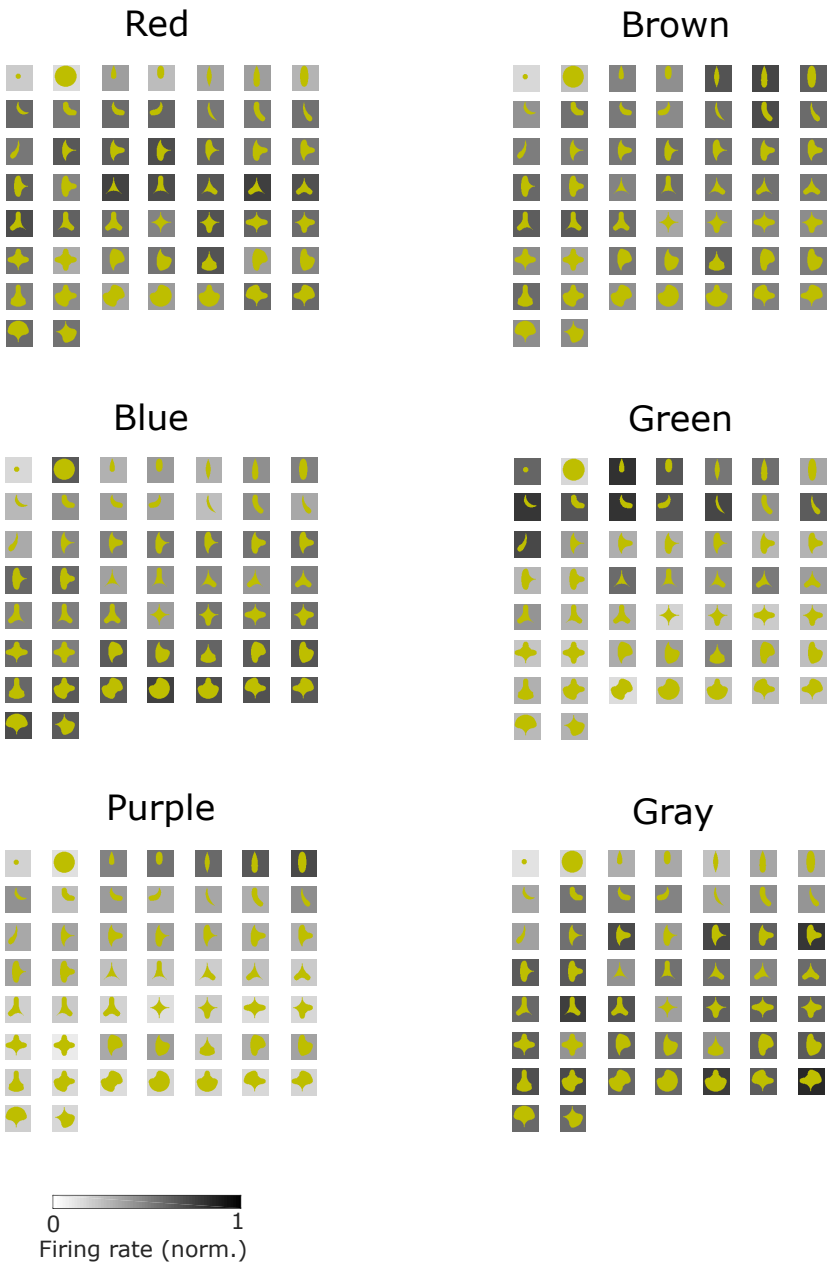
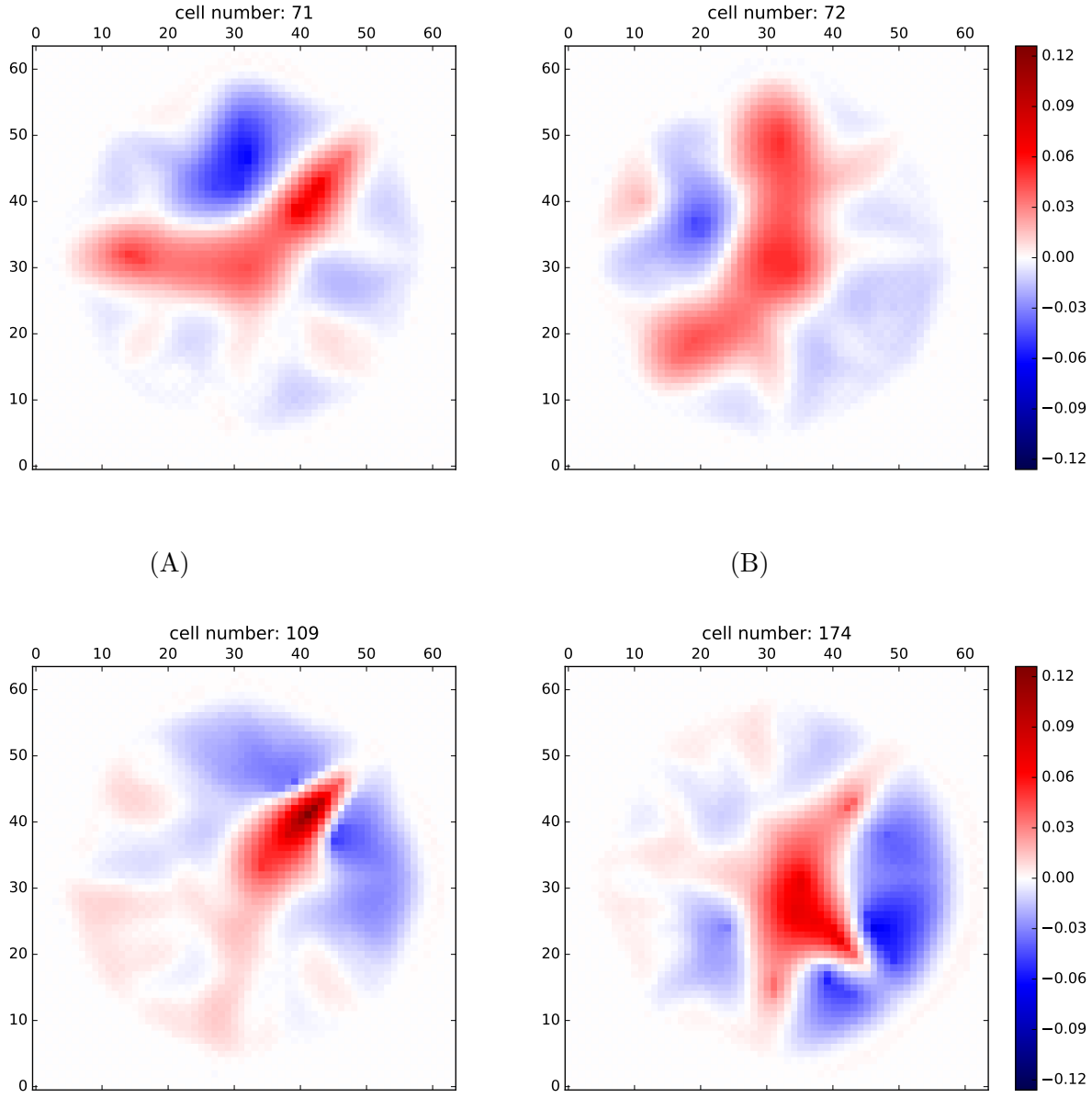


Figure 4.14: The average normalized response of cells in each cluster to each distinct shape. This response is calculated by taking the maximum response over all the 8 rotations of a shape.



(A) (B)

(C) (D)

Figure 4.15: Linear filters from LN fits for cells #71 and #72 in the brown cluster and cells #109 and #174 in the red cluster. The r-value for the LN models is .66 for cells #72, #109, #174 and .75 for cell #71.

BIBLIOGRAPHY

[AB03] Charalambos D Aliprantis and Owen Burkinshaw. *Locally solid Riesz spaces with applications to economics*. Number 105. American Mathematical Soc., 2003.

[ACP14] Yossi Azar, Ilan Reuven Cohen, and Debmalya Panigrahi. Online covering with convex objectives and applications. *arXiv preprint arXiv:1412.3507*, 2014.

[AD14] Shipra Agrawal and Nikhil R Devanur. Fast algorithms for online stochastic convex programming. *arXiv preprint arXiv:1410.7596*, 2014.

[AHR08] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, pages 263–274, 2008.

[AJG08] Fabrice Arcizet, Christophe Jouffrais, and Pascal Girard. Natural textures classification in area v4 of the macaque monkey. *Experimental brain research*, 189(1):109–120, 2008.

[AJG09] Fabrice Arcizet, Christophe Jouffrais, and Pascal Girard. Coding of shape from shading in area v4 of the macaque monkey. *BMC neuroscience*, 10(1):140, 2009.

[ANW10] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. pages 37–45, 2010.

[AS04] Alexander A Ageev and Maxim I Sviridenko. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *Journal of Combinatorial Optimization*, 8(3):307–328, 2004.

[AWY09] Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *arXiv preprint arXiv:0911.2974*, 2009.

[BB17] Jeffrey Bilmes and Wenruo Bai. Deep submodular functions. *arXiv preprint arXiv:1701.08939*, 2017.

[BCG+14] Niv Buchbinder, Shahar Chen, Anupam Gupta, Viswanath Nagarajan, et al. Online packing and covering framework with convex objectives. *arXiv preprint arXiv:1412.8347*, 2014.

[Ber75] Dimitri P Bertsekas. Necessary and sufficient conditions for a penalty method to be exact. *Mathematical programming*, 9(1):87–99, 1975.

[Ber14] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[BIKK08] Moshe Babaioff, Nicole Immorlica, David Kempe, and Robert Kleinberg. Online auctions and generalized secretary problems. *SIGecom Exch.*, 7(2):7:1–7:11, June 2008.

[BJN07] Niv Buchbinder, Kamal Jain, and Joseph Seffi Naor. Online primal-dual algorithms for maximizing ad-auctions revenue. In *Algorithms–ESA 2007*, pages 253–264. Springer, 2007.

[BN09] Niv Buchbinder and Joseph Naor. Online primal-dual algorithms for covering and packing. *Mathematics of Operations Research*, 34(2):270–286, 2009.

[BP12] Brittany N Bushnell and Anitha Pasupathy. Shape encoding consistency across colors in primate v4. *Journal of neurophysiology*, 108(5):1299–1308, 2012.

[BTW+07] Florentina Bunea, Alexandre Tsybakov, Marten Wegkamp, et al. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[Bur91] James V Burke. An exact penalization viewpoint of constrained optimization. *SIAM Journal on control and optimization*, 29(4):968–998, 1991.

[CC84] Michele Conforti and Gérard Cornuéjols. Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the rado-edmonds theorem. *Discrete applied mathematics*, 7(3):251–274, 1984.

[CCPV07] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a submodular set function subject to a matroid constraint. In *IPCO*, volume 7, pages 182–196. Springer, 2007.

[CCS10] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[CHK15] TH Chan, Zhiyi Huang, and Ning Kang. Online convex covering and packing problems. *arXiv preprint arXiv:1502.01802*, 2015.

[CKP+07] Charles Cadieu, Minjoon Kouh, Anitha Pasupathy, Charles E Connor, Maximilian Riesenhuber, and Tomaso Poggio. A model of v4 shape selectivity and invariance. *Journal of neurophysiology*, 98(3):1733–1750, 2007.

[Cla90] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[CP11] E.J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

[CR12] Emmanuel Candès and Benjamin Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, pages 1–13, 2012.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[CRT06] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[CT06] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

[CT07] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[DH09] Nikhil R. Devanur and Thomas P. Hayes. The adwords problem: Online keyword matching with budgeted bidders under random permutations. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 71–78, New York, NY, USA, 2009. ACM.

[DHG06] Stephen V David, Benjamin Y Hayden, and Jack L Gallant. Spectral receptive field properties explain shape selectivity in area v4. *Journal of neurophysiology*, 96(6):3492–3505, 2006.

[DJ12] Nikhil R Devanur and Kamal Jain. Online matching with concave returns. In *Proceedings of the 44th symposium on Theory of Computing*, pages 137–144. ACM, 2012.

[DJSW11] Nikhil R Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 29–38. ACM, 2011.

[DL16] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *arXiv preprint arXiv:1602.06661*, 2016.

[Don06] David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[Dra03] Sever S Dragomir. *Some Gronwall type inequalities and applications*. Nova Science, 2003.

[DS06] Shahar Dobzinski and Michael Schapira. An improved approximation algorithm for combinatorial auctions with submodular bidders. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1064–1073. Society for Industrial and Applied Mathematics, 2006.

[EKS+96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[ESF14] Reza Eghbali, Jon Swenson, and Maryam Fazel. Exponentiated subgradient algorithm for online optimization under the random permutation model. *arXiv preprint arXiv:1410.7171*, 2014.

[FHK+10] Jon Feldman, Monika Henzinger, Nitish Korula, Vahab S. Mirrokni, and Cliff Stein. Online stochastic packing applied to display ad allocation. In *Proceedings of the 18th Annual European Conference on Algorithms: Part I*, ESA'10, pages 182–194, Berlin, Heidelberg, 2010. Springer-Verlag.

[FMMM09] Jon Feldman, Aranyak Mehta, Vahab Mirrokni, and S Muthukrishnan. Online stochastic matching: Beating 1-1/e. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, pages 117–126. IEEE, 2009.

[Fuj05] Satoru Fujishige. *Submodular functions and optimization*, volume 58. Elsevier, 2005.

[FW56] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[GBVE93] Jack L Gallant, Jochen Braun, and David C Van Essen. Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, 259(5091):100–103, 1993.

[GM14] Anupam Gupta and Marco Molinaro. How the experts algorithm can help solve lps online. *arXiv preprint arXiv:1407.5298*, 2014.

[Gor85] Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[Gro11] David Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

[GS07] Pranava R Goundan and Andreas S Schulz. Revisiting the greedy approach to submodular set function maximization. *Optimization online*, pages 1–25, 2007.

[Han13] Frank Hansen. The fast track to löwners theorem. *Linear Algebra and its Applications*, 438(11):4557–4571, 2013.

[HYZ08] Elaine T Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for \ell_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

[HZSL13] Ke Hou, Zirui Zhou, Anthony M So, and Zhi-quan Luo. On the linear convergence of the proximal gradient method for trace norm regularization. In *Advances in Neural Information Processing Systems*, pages 710–718, 2013.

[JB09] Siddharth Joshi and Stephen Boyd. Sensor Selection via Convex Optimization. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 57(2), 2009.

[JL12] Patrick Jaillet and Xin Lu. Near-optimal online algorithms for dynamic resource allocation problems. *arXiv preprint arXiv:1208.2596*, 2012.

[JL13] Patrick Jaillet and Xin Lu. Online stochastic matching: New algorithms with better bounds. *Mathematics of Operations Research*, 2013.

[JMD10] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *NIPS*, volume 23, pages 937–945, 2010.

[JYZ13] Rong Jin, Tianbao Yang, and Shenghuo Zhu. A new analysis of compressive sensing by stochastic proximal gradient descent. *CoRR*, abs/1304.4680, 2013.

[KESFP14] Yoshito Kosai, Yasmine El-Shamayleh, Amber M Fyall, and Anitha Pasupathy. The role of visual area v4 in the discrimination of partially occluded shapes. *The Journal of Neuroscience*, 34(25):8570–8584, 2014.

[Kle05] Robert Kleinberg. A multiple-choice secretary algorithm with applications to online auctions. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 630–631. Society for Industrial and Applied Mathematics, 2005.

[KMT11] Chinmay Karande, Aranyak Mehta, and Pushkar Tripathi. Online bipartite matching with unknown distributions. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 587–596. ACM, 2011.

[KP00] Bala Kalyanasundaram and Kirk R Pruhs. An optimal deterministic algorithm for online b-matching. *Theoretical Computer Science*, 233(1):319–325, 2000.

[KPV] M. Kapralov, I. Post, and J. Vondrák. Online submodular welfare maximization: Greedy is optimal. In *Proceedings of the Twenty-Fourth Annual Symposium on Discrete Algorithms*.

[KRTV14] Thomas Kesselheim, Klaus Radke, Andreas Tönnis, and Berthold Vöcking. Primal beats dual on online packing lps in the random-order model. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 303–312, New York, NY, USA, 2014. ACM.

[KT94] Eucaly Kobatake and Keiji Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of neurophysiology*, 71(3):856–867, 1994.

[KVV90] Richard M Karp, Umesh V Vazirani, and Vijay V Vazirani. An optimal algorithm for on-line bipartite matching. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 352–358. ACM, 1990.

[Lor53] GG Lorentz. An inequality for rearrangements. *The American Mathematical Monthly*, 60(3):176–179, 1953.

[LPVDG+11] Karim Lounici, Massimiliano Pontil, Sara Van De Geer, Alexandre B Tsybakov, et al. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

[LT92] Zhi-Quan Luo and Paul Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

[LT13] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 2013.

[LV09] Zhang Liu and Lieven Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[LX13] Zhaosong Lu and Lin Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, pages 1–28, 2013.

[MGC11] S. Ma, D. Goldfarb, and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1):321–353, 2011.

[MGS12] Vahideh H Manshadi, Shayan Oveis Gharan, and Amin Saberi. Online stochastic matching: Online actions based on offline statistics. *Mathematics of Operations Research*, 37(4):559–573, 2012.

[MHT10] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

[Mor62] Jean-Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien.(french). *CR Acad. Sci. Paris*, 255:2897–2899, 1962.

[MPTJ07] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

[MSVV07] Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22, 2007.

[MY11] Mohammad Mahdian and Qiqi Yan. Online bipartite matching with random arrivals: an approach based on strongly factor-revealing lps. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 597–606. ACM, 2011.

[Nes05] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[Nes09] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[Nes12] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[Nes13] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[NHNT13] Sang (Peter) Chin Nam H. Nguyen and Trac D. Tran. A unified iterative greedy algorithm for sparsity-constrained optimization. *Submitted*, 2013.

[NN04] Yurii Nesterov and IU E Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

[NN13] Yurii Nesterov and Arkadi Nemirovski. On first-order algorithms for l 1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

[NNW14] Nam Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *arXiv preprint arXiv:1407.0088*, 2014.

[NRWY12] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[NT09] Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[NWF78] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

[ONP16] Timothy D. Oleskiw, Amy Nowack, and Anitha Pasupathy. Joint coding of shape and blur in area v4: toward a sufficient representation of natural scenes. *COSYNE*, 2016.

[OPB14] Timothy D Oleskiw, Anitha Pasupathy, and Wyeth Bair. Spectral receptive fields do not explain tuning for boundary curvature in v4. *Journal of neurophysiology*, 112(9):2114–2122, 2014.

[PBP] Dina Popovkina, Wyeth Bair, and Anitha Pasupathy. Advancing models of shape representation for mid-level vision. *In prep.*

[PC01] Anitha Pasupathy and Charles E Connor. Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*, 86(5):2505–2519, 2001.

[Pol79] BT Poljak. On the bertsekas method for minimization of composite functions. In *International Symposium on Systems Optimization and Analysis*, pages 179–186. Springer, 1979.

[Puk93] Friedrich Pukelsheim. *Optimal design of experiments*, volume 50. siam, 1993.

[Rei43] Anders Reiz. On the numerical solution of certain types of integral equations. *Meddelanden fran Lunds Astronomiska Observatorium Serie I*, 161:1–21, 1943.

[RFP10] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[Roc76] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[RT+11] Angelika Rohde, Alexandre B Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[RT14] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

[RWW98] R Tyrrell Rockafellar, Roger J-B Wets, and Maria Wets. *Variational analysis*, volume 317. Springer, 1998.

[RWY11] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[SBV10] Manohar Shamaiah, Siddhartha Banerjee, and Haris Vikalo. Greedy sensor selection: Leveraging submodularity. In *49th IEEE Conference on Decision and Control (CDC)*, pages 2572–2577. IEEE, dec 2010.

[Sch02] Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer Science & Business Media, 2002.

[Sim14] Dan A Simovici. On submodular and supermodular functions on lattices and related structures. In *Multiple-Valued Logic (ISMVL), 2014 IEEE 44th International Symposium on*, pages 202–207. IEEE, 2014.

[SS11] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[SSGS11] Shai Shalev-Shwartz, Alon Gonen, and Ohad Shamir. Large-scale convex minimization with a low-rank constraint. *arXiv preprint arXiv:1106.1622*, 2011.

[SSS07a] Shai Shalev-Shwartz and Yoram Singer. Online learning: Theory, algorithms, and applications. 2007.

[SSS07b] Shai Shalev-Shwartz and Yoram Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2-3):115–142, 2007.

[SSSZ10] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

[Svi04] Maxim Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 32(1):41–43, 2004.

[SW97] Paola Sebastiani and Henry P Wynn. Bayesian experimental design and shannon information. In *Proceedings of the Section on Bayesian Statistical Science*, volume 44, pages 176–181. The Association, 1997.

[Tal05] Michel Talagrand. *The generic chaining*, volume 154. Springer, 2005.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Top11] Donald M Topkis. *Supermodularity and complementarity*. Princeton university press, 2011.

[TY10] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

[VDGB+09] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[Von08] Jan Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 67–74. ACM, 2008.

[Von10] Jan Vondrák. Submodularity and curvature: the optimal algorithm. *RIMS Kokyuroku Bessatsu B*, 23:253–266, 2010.

[WNF09] Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *Signal Processing, IEEE Transactions on*, 57(7):2479–2493, 2009.

[WYGZ10] Zaiwen Wen, Wotao Yin, Donald Goldfarb, and Yin Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM Journal on Scientific Computing*, 32(4):1832–1857, 2010.

[XZ13] Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

[Zek83] Semir Zeki. The distribution of wavelength and orientation selective cells in different areas of monkey visual cortex. *Proceedings of the Royal Society of London B: Biological Sciences*, 217(1209):449–470, 1983.

[ZJL13] Haibin Zhang, Jiaojiao Jiang, and Zhi-Quan Luo. On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. *Journal of the Operations Research Society of China*, 1(2):163–186, 2013.

[ZP08] Michael M Zavlanos and George J Pappas. Distributed connectivity control of mobile networks. *Robotics, IEEE Transactions on*, 24(6):1416–1428, 2008.

[ZS17] Zirui Zhou and Anthony Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, pages 1–40, 2017.

Appendix A

SAMPLE COMPLEXITY FOR ASSUMPTION 1

We give a lower bound on the number of measurements m that suffice for the existence of $r > 1$ in Assumption 1 with high probability when A is sampled from a certain class of distributions. To simplify the notation we assume that $B = I$; therefore, $\langle x, y\rangle = x^T y$. Given a random variable z the sub-Gaussian norm of z is defined as:

$$\|z\|_{\psi_2} = \inf\{\beta > 0 \,|\, \mathbb{E}\psi_2\left(\frac{|z|}{\beta}\right) \leq 1\},$$

where $\psi_2(x) = e^{x^2} - 1$. For an n dimensional random vector $w \sim P$ the sub-Gaussian norm is defined as

$$\|w\|_{\psi_2} = \sup_{u \in S^{n-1}} \|\langle w, u\rangle\|_{\psi_2}.$$

P is called isotropic if $E\left[\langle w, u\rangle^2\right] = 1$ for all $u \in S^{n-1}$. Two important examples of sub-Gaussian random variables are Gaussian and bounded random variables. Suppose $A : \mathbb{R}^n \mapsto \mathbb{R}^m$ is given by:

$$(Ax)_i = \frac{1}{\sqrt{m}}\langle A_i, x\rangle \quad \forall i \in \{1, 2, \ldots, m\}, \tag{A.1}$$

where A_i, $1 \leq i \leq m$ are iid samples from an isotropic sub-Gaussian distribution P on $\mathbb{R}^n$. Two important examples are standard Gaussian vector $A_i \sim \mathcal{N}(0, I_n)$ and random vector of independent Rademacher variables [1]. We want to bound the following probabilities for $\theta \in (0, 1)$:

$$P(\rho_-(A, k) < 1 - \theta) \tag{A.2}$$

$$P(\rho_+(A, k) > 1 + \theta). \tag{A.3}$$

[1]For general psd B, the example are $A_i = B^{-\frac{1}{2}} A_i'$ with $A_i' \sim \mathcal{N}(0, I_n)$ or $A_{i,j}'$ Rademacher for all j.

When $A_i \sim \mathcal{N}(0, I_n)$ for all i, one can use the generalization of Slepian's lemma by Gordon [Gor85] alongside concentration inequalities for Lipschitz function of Gaussian random variable to derive (see, for example, [LT13, chapter 15]):

$$P(\sqrt{\rho_-(A,k)} < \sqrt{\frac{m}{m+1}} - \theta) \leq e^{-\frac{m\theta^2}{8}},$$
$$P(\sqrt{\rho_+(A,k)} > 1 + \theta) \leq e^{-\frac{m\theta^2}{8}},$$

whenever,

$$\theta \geq \frac{2G(k)}{\sqrt{m}}.$$

Here, G is defined as:

$$G(k) := \mathbb{E} \sup_{u \in \sqrt{k}\mathcal{B}_{\|\cdot\|} \cap S^{n-1}} |\langle u, g \rangle|,$$

where $g \sim \mathcal{N}(0, I_n)$. For sub-Gaussian case, we use a result by Mendelson et al.[MPTJ07, Theorem 2.3]. Using Talgrand's generic chaining theorem [Tal05, Theorem 2.1.1], the authors have given a result, which similar to the Gaussian case depends on $G(k)$. Their result in our notation states:

Proposition 7. *Suppose A is given by* (A.1)*. If P is an isotropic distribution and $\|A_1\|_{\psi_2} \leq \alpha$, then there exist constants c_1 and c_2 such that*

$$\rho_-(A/\sqrt{m}, k) \geq 1 - \theta, \tag{A.4}$$

$$\rho_+(A/\sqrt{m}, k) \leq 1 + \theta, \tag{A.5}$$

with probability exceeding $1 - \exp(-c_2\theta^2 m/\alpha^4)$ whenever

$$\theta \geq \frac{c_1\alpha^2 G(k)}{\sqrt{m}}.$$

Suppose $\lambda_{\text{tgt}} = 4\|A^* z\|^*$, which sets $\gamma = \frac{5+4\delta}{3-4\delta}$. We can state the following proposition based on Proposition 7 :

Proposition 8. *Let* $r > 1$, $\tilde{k} = 36rck_0(1+\gamma)\gamma_{\text{inc}}$ *and* $\bar{k} = ck_0(1+\gamma)^2$. *If* $m \geq \frac{c_1\alpha^4}{(r-1)^2}(G(2\tilde{k})^2 + r^2G(\bar{k})^2)$, *then* r *satisfies Assumption 1 with probability exceeding* $1-\exp(c_2(r-1)^2m/r^2\alpha^2)$.

The proof is a simple adaptation of proof of Theorem 1.4 in [MPTJ07] which we omit here. To compare this with the number of measurements sufficient for successful recovery within a given accuracy, by combining (B.20) in the proof Lemma 1 and Proposition 7 we get:

Proposition 9. *Let* $r > 1$, $\bar{k} = ck_0(1+\gamma)^2$ *and* $x^* \in \operatorname{argmin}\phi_\lambda(x)$. *If* $m \geq \frac{c_1\alpha^4r^2}{(r-1)^2}G(\bar{k})^2$, *then* $\|x^* - x_0\|_2 \leq c_2r\lambda\sqrt{ck_0}$ *with probability exceeding* $1-\exp(c_2(r-1)^2m/r^2\alpha^2)$.

Note that this bound on m in case of l_1, $l_{1,2}$ and nuclear norms orderwise matches the lower bounds given by minimax rates in [RWY11], [LPVDG$^+$11] and [RT$^+$11].

Appendix B

PROOFS FROM CHAPTER 2

B.1 Proof of Theorem 1

Sufficiency. First consider the case where $k = 1$ and $x = \gamma_1 a_1$ with $\gamma_1 > 0$. Note that $a_1 \in \partial\|x\| = \partial\|a_1\|$ because $\|a_1\|^* = 1$ for all $a_1 \in \mathcal{G}_{\|\cdot\|}$ and $\langle a_1, x\rangle = \gamma_1 = \|x\|$. Define:

$$C = \{\xi - a_1 | \xi \in \partial\|a_1\|\}.$$

Note that C is a convex set that contains the origin. Moreover, C is orthogonal to a_1. We claim that (2.9) is satisfied with $T_{a_1}^{\perp} = \operatorname{span} C$. To establish the claim, we first prove that C is symmetric and is contained in the dual norm ball. Let $v \in C$ and $\xi = a_1 + v \in \partial\|a_1\|$. By (2.4), $\langle a_1, \xi\rangle = \|\xi\|^* = 1$. Therefore,

$$a_1 \in \underset{a \in \mathcal{G}_{\|\cdot\|}}{\operatorname{argmax}} \langle a, \xi\rangle$$

and we can apply the hypothesis of the theorem (in particular statement I) to obtain an orthonormal representation for ξ:

$$\xi = a_1 + \sum_{i=1}^{l} \eta_i b_i.$$

Now by statement II in the hypothesis we get:

$$\|v\|^* = \max_i \eta_i \leq \|\xi\|^* \leq 1.$$

Let $\xi' = a_1 - \sum_{i=1}^{l} \eta_i b_i$. By the hypothesis, $\|\xi'\|^* = \max\{1, \max_i \eta_i\} = 1$. Also, $\langle \xi', a_1\rangle = 1$ hence $\xi' \in \partial\|a_1\|$ and $-v \in C$.

Let $v \in \operatorname{span} C$ with $\|v\|^* \leq 1$. Since C is a symmetric convex set, there exists $\lambda \in (0, 1]$ such that $\lambda v \in C$ (i.e., C is absorbing in $\operatorname{span} C$). Define $z = a_1 + \lambda v$ which is in $\partial\|a_1\|$. Since $\langle a_1, z\rangle = \|z\|^* = 1$, we can write z as

$$z = a_1 + \sum_{i=1}^{k'} \nu_i c_i,$$

where $\{c_i | i = 1, \ldots, k'\} \subset \mathcal{G}_{\|\cdot\|}$ and $\{\nu_i \geq 0 | i = 1, \ldots, k'\}$ satisfy the hypothesis of the theorem. In particular, since $v = 1/\lambda \sum_{i=1}^{k'} \nu_i c_i$, we have $\max_i \nu_i/\lambda \leq 1$. Hence $\|a_1 + v\|^* = \max\{1, \nu_1/\lambda, \ldots \nu_{t'}/\lambda\} = 1$ and $a_1 + v \in \partial\|a_1\|$. Therefore,

$$\partial\|a_1\| = \{a_1 + v | v \in \operatorname{span} C, \|v\|^* \leq 1\}.$$

Now suppose that $x = \sum_{i=1}^{k} \gamma_i a_i$ with $k > 1$. Note that $\sum_{i=1}^{k} a_i \in \partial\|x\|$ since $\left\|\sum_{i=1}^{k} a_i\right\|^* = 1$ and $\langle \sum_{i=1}^{k} a_i, x\rangle = \sum_{i=1}^{k} \gamma_i = \|x\|$. Let $\xi \in \partial\|x\|$ and define $v = \xi - \sum_{i=1}^{k} a_i$. We can write:

$$\|x\| = \sum_{i=1}^{k} \gamma_i = \langle \xi, x\rangle = \sum_{i=1}^{k} \gamma_i \langle \xi, a_i\rangle$$

$$\Rightarrow \forall i \in \{1, 2, \ldots, k\} : \langle \xi, a_i\rangle = 1 \Rightarrow \forall i \in \{1, 2, \ldots, k\} : \xi \in \partial\|a_i\|. \tag{B.1}$$

Also, since $\sum_{i=1}^{k} a_i \in \partial\|a_i\|$, (B.1) results in:

$$\forall i \in \{1, 2, \ldots, k\} : \quad v \in T_{a_i}^{\perp}. \tag{B.2}$$

Since $\xi = \sum_{i=1}^{k} a_i + v \in \partial\|a_1\|$, we have $\left\|\sum_{i=2}^{k} a_i + v\right\|^* = 1$ hence $\sum_{i=2}^{k} a_i + v \in \partial\|a_2\|$. By induction, we conclude that $a_k + v \in \partial\|a_k\|$. This implies $\|v\|^* \leq 1$.

Let $v' \in \cap_{i\in\{1,2,\ldots,k\}} T_{a_i}^{\perp}$ with $\|v'\|^* \leq 1$ and define $\xi' = \sum_{i=1}^{k} a_i + v'$. We will prove that $\|\xi'\|^* \leq 1$ and hence $\xi' \in \partial\|x\|$. To prove this we use induction. Define

$$z_l = \sum_{i=k-l+1}^{k} a_i + v' \qquad \forall l \in \{1, 2, \ldots, k\}.$$

Note that $\|z_1\|^* \leq 1$ since $z_1 = a_k + v' \in \partial\|a_k\|$. Suppose $\|z_{l'}\|^* \leq 1$ for some $l' < k$. We prove that $\|z_{l'+1}\|^* \leq 1$. We have $\sum_{i=k-l'+1}^{k} a_i \in T_{a_{k-l'}}^{\perp}$ because $\sum_{i=k-l'}^{k} a_i = a_{k-l'} + \sum_{i=k-l'+1}^{k} a_i \in \partial\|a_{k-l'}\|$. Combining this with the fact that $v' \in T_{a_{k-l'}}^{\perp}$, we get $z_{l'} \in T_{a_{k-l'}}^{\perp}$. Therefore,

$z_{l'+1} = a_{k-l'} + z_{l'} \in \partial\|a_{k-l'}\|$ hence $\|z_{l'+1}\|^* \leq 1$. Thus $\|\xi'\|^* = \|z_k\|^* \leq 1$. We conclude that:

$$\partial\|x\| = \{\sum_{i=1}^{k} a_i + v | v \in \bigcap_{i=1}^{k} T_{a_i}^{\perp}, \|v\|^* \leq 1\}. \tag{B.3}$$

Necessity. For any $a \in \mathcal{G}_{\|\cdot\|}$, we have:

$$\langle a, a\rangle = 1,$$
$$\forall b \in \mathcal{G}_{\|\cdot\|}: \quad \langle b, a\rangle \leq \|b\|_2 \|a\|_2 = 1.$$

That implies $\|a\|^* = 1$ and $a \in \partial\|a\|$. Since $a \in T_a$, we conclude that:

$$\partial\|a\| = \{a + v | v \in T_a^{\perp}, \|v\|^* \leq 1\}. \tag{B.4}$$

Take $\gamma_1 = \langle a_1, x\rangle = \|x\|^*$ and let $\Delta_1 = x - \gamma_1 a_1$. If $\Delta_1 = 0$, then take $k = 1$ and $x = \gamma_1 a_1$. Suppose $\Delta_1 \neq 0$. Since $\left\|\frac{1}{\gamma_1}x\right\|^* = 1$ and $\langle a_1, \frac{1}{\gamma_1}x\rangle = \|a_1\| = 1$, we can conclude that $\frac{1}{\gamma_1}x \in \partial\|a_1\|$. Furthermore, we have

$$P_{T_{a_1}{}^{\perp}}(x) = x - \gamma_1 P_{T_{a_1}}(\frac{1}{\gamma_1}x) = x - \gamma_1 a_1 = \Delta_1 \tag{B.5}$$
$$\Rightarrow \Delta_1 \in T_{a_1}^{\perp}.$$

Now we introduce a lemma that will be used in the rest of the proof.

Lemma 7. *Suppose $a \in \mathcal{G}_{\|\cdot\|}$ and $y \in T_a^{\perp} - \{0\}$. If $z \in \mathcal{B}_{\|\cdot\|}$ is such that $\|y\|^* = \langle y, z\rangle$, then $z \in T_a^{\perp}$.*

Proof. Without loss of generality assume that $\|y\|^* = 1$. It suffices to show that if $b \in \mathcal{G}_{\|\cdot\|}$ and $\langle y, b\rangle = 1$, then $b \in T_a^{\perp}$. Consider such $b \in \mathcal{G}_{\|\cdot\|}$. By (B.4), $\|a + y\|^* = 1$. That results in:

$$1 \geq \langle a + y, b\rangle = \langle a, b\rangle + 1 \Rightarrow 0 \geq \langle a, b\rangle.$$

By considering $-y$ and $-b$ we get that $\langle a, b\rangle = 0$. Since $\langle a + y, b\rangle = \|b\| = 1$, we can conclude that $a + y \in \partial\|b\|$. Since $\langle y, b\rangle = 1$ and $\|y\|^* = 1$, $y \in \partial\|b\|$. Combining these two conclusions, we get:

$$y \in \partial\|b\|, a+y \in \partial\|b\| \Rightarrow a \in T_b^{\perp} \overset{(B.4)}{\Longrightarrow} \|a+b\|^* \leq 1 \Rightarrow a+b \in \partial\|a\| \Rightarrow b \in T_a^{\perp}$$

□ □

Suppose that there exist $l \in \{1,2,\ldots,k\}$, an orthogonal set $\{a_i \in \mathcal{G}_{\|\cdot\|} |\, i=1,2,\ldots,l\}$, and a set of coefficients $\{\gamma_i \geq 0 |\, i=1,2,\ldots,l\}$ such that $x = \sum_{i=1}^{l} \gamma_i a_i + \Delta_l$, $\Delta_l \in \cap_{i=1}^{l} T_{a_i}^{\perp}$, and:

$$\partial\left\|\sum_{i=1}^{l} a_i\right\| = \{\sum_{i=1}^{l} a_i + v | v \in \bigcap_{i=1}^{l} T_{a_i}^{\perp}, \|v\|^* \leq 1\}. \tag{B.6}$$

By Lemma 7, there exists $a_{l+1} \in \mathcal{G}_{\|\cdot\|}$ such that $a_{l+1} \in \cap_{i=1}^{l} T_{a_i}^{\perp}$ and $\langle a_{l+1}, \Delta_l \rangle = \|\Delta_l\|^*$. Take $\gamma_{l+1} = \langle a_{l+1}, \Delta_l \rangle = \|\Delta_l\|^*$ and let $\Delta_{l+1} = \Delta_l - \gamma_{l+1} a_{l+1}$. We have $\Delta_{l+1} \in \bigcap_{i=1}^{l} T_{a_i}^{\perp}$ because $\{\Delta_l, a_{l+1}\} \subset \bigcap_{i=1}^{l} T_{a_i}^{\perp}$. Since $\left\|\frac{1}{\gamma_{l+1}}\Delta_l\right\|^* = 1$ and $\langle a_{l+1}, \frac{1}{\gamma_{l+1}}\Delta_l \rangle = \|a_{l+1}\| = 1$, we can conclude that $\frac{1}{\gamma_{l+1}}\Delta_l \in \partial\|a_{l+1}\|$. Using the same reasoning as in (B.5), we have $\Delta_{l+1} \in T_{a_{l+1}}^{\perp}$ hence $\Delta_{l+1} \in \bigcap_{i=1}^{l+1} T_{a_i}^{\perp}$.

By decomposability assumption there exists $e \in \mathbb{R}^n$ and a subspace T such that:

$$\partial\left\|\sum_{i=1}^{l+1} a_i\right\| = \{e + v | v \in T^{\perp}, \|v\|^* \leq 1\}. \tag{B.7}$$

We claim that

$$e = \sum_{i=1}^{l+1} a_i \tag{B.8}$$

$$T^{\perp} = \bigcap_{i=1}^{l+1} T_{a_i}^{\perp}. \tag{B.9}$$

To prove the first claim, it is enough to show that $\sum_{i=1}^{l+1} a_i \in \partial\left\|\sum_{i=1}^{l+1} a_i\right\|$. Note that $\left\|\sum_{i=1}^{l+1} a_i\right\|^* \leq 1$ since $\sum_{i=1}^{l+1} a_i = \sum_{i=1}^{l} a_i + a_{l+1} \in \partial\left\|\sum_{i=1}^{l} a_i\right\|$ which is given by (B.6). Now we can write:

$$l+1 = \langle \sum_{i=1}^{l+1} a_i, \sum_{i=1}^{l+1} a_i \rangle \leq \left\| \sum_{i=1}^{l+1} a_i \right\| \left\| \sum_{i=1}^{l+1} a_i \right\|^* \leq \left\| \sum_{i=1}^{l+1} a_i \right\|.$$

On the other hand, by triangle inequality,

$$\left\| \sum_{i=1}^{l+1} a_i \right\| \leq \left\| \sum_{i=1}^{l} a_i \right\| + \|a_{l+1}\| = l+1,$$

thus

$$\left\| \sum_{i=1}^{l+1} a_i \right\| = \langle \sum_{i=1}^{l+1} a_i, \sum_{i=1}^{l+1} a_i \rangle.$$

Therefore, $\sum_{i=1}^{l+1} a_i \in \partial \left\| \sum_{i=1}^{l+1} a_i \right\|$. Since $\sum_{i=1}^{l+1} a_i \in T_{\sum_{i=1}^{l+1} a_i} = T$, we conclude that:

$$\partial \left\| \sum_{i=1}^{l+1} a_i \right\| = \{ \sum_{i=1}^{l+1} a_i + v | v \in T^\perp, \|v\|^* \leq 1 \}.$$

To prove (B.9), we first show that $\bigcap_{i=1}^{l+1} T_{a_i}^\perp \in T^\perp$. Let $\xi = e + v$ with $v \in \bigcap_{i=1}^{l+1} T_{a_i}^\perp$. Note that $\|a_{l+1} + v\|^* \leq 1$ since $a_{l+1} + v \in \partial \|a_{l+1}\|$. Furthermore, $a_{l+1} + v \in \bigcap_{i=1}^{l} T_{a_i}^\perp$, which in turn implies $\sum_{i=1}^{l+1} a_i + v \in \partial \left\| \sum_{i=1}^{l} a_i \right\|$ hence $\left\| \sum_{i=1}^{l+1} a_i + v \right\|^* \leq 1$. Additionally, we have:

$$\langle \sum_{i=1}^{l+1} a_i + v, \sum_{i=1}^{l+1} a_i \rangle = \left\| \sum_{i=1}^{l+1} a_i \right\| = l+1.$$

Hence $\xi \in \partial \left\| \sum_{i=1}^{l+1} a_i \right\|$ and $v \in T^\perp$.

Now, let $\xi' = \sum_{i=1}^{l+1} a_i + v' \in \left\| \sum_{i=1}^{l+1} a_i \right\|$. Note that:

$$\langle \xi', \sum_{i=1}^{l+1} a_i \rangle = \langle \xi', \sum_{i=1}^{l} a_i \rangle + \langle \xi', a_{l+1} \rangle = l+1 \Rightarrow \langle \xi', \sum_{i=1}^{l} a_i \rangle = l, \langle \xi', a_{l+1} \rangle = 1 \Rightarrow \xi' \in \partial \left\| \sum_{i=1}^{l} a_i \right\| \cap \partial \|a_{l+1}\|$$

$$\Rightarrow v' \in \bigcap_{i=1}^{l} T_{a_i}^\perp, \sum_{i=1}^{l} a_i + v' \in \bigcap_{i=1}^{l} T_{a_i}^\perp;$$

moreover, $\sum_{i=1}^{l} a_i \in T_{a_{l+1}}^\perp$ since $\sum_{i=1}^{l+1} a_i \in \partial \|a_{l+1}\|$. This implies $v \in \bigcap_{i=1}^{l+1} T_{a_i}^\perp$ which completes the proof of (B.9).

Because $a_i \notin T_{a_i}^{\perp}$ for all $i \in \{1, 2, \ldots, l+1\}$, $dim(\cap_{i=1}^{l+1} T_{a_i}^{\perp}) \leq n-l-1$. Hence there exists $k \leq n$, an orthogonal set $\{a_i \in \mathcal{G}_{\|\cdot\|},\ i = 1, 2, \ldots, k\}$, and a set of coefficients $\{\gamma_i \geq 0, i \in \{1, 2, \ldots, k\}\}$ such that $x = \sum_{i=1}^{k} \gamma_i a_i$ and:

$$\partial \left\| \sum_{i=1}^{k} a_i \right\| = \{ \sum_{i=1}^{k} a_i + v \mid v \in \bigcap_{i=1}^{k} T_{a_i}^{\perp}, \|v\|^* \leq 1 \}. \tag{B.10}$$

That proves $\|x\| = \langle \sum_{i=1}^{k} a_i, x \rangle = \sum_{i=1}^{k} \gamma_i$.

To prove statement II, we first prove that $a_i \in T_{a_j}^{\perp}$ for all $i, j \in \{1, 2, \ldots, k\}$. By (B.10), $\left\| \sum_{i=1}^{k} a_i \right\|^* \leq 1$. We can write:

$$\langle \sum_{i=1}^{k} a_i, a_j \rangle = 1 \Rightarrow \sum_{i=1}^{k} a_i \in \partial \|a_j\| \Rightarrow \sum_{i=1}^{k} a_i - a_j \in T_{a_j}^{\perp},\ \left\| \sum_{i=1}^{k} a_i - a_j \right\|^* \leq 1,$$

Now the claim follows from Lemma 7.

Let $l = |\{\eta_i | \eta_i \neq 0\}|$. If $l = 0$, the statement is trivially true. Suppose the statement is true when $l = l' - 1$ for some $l' \in \{1, \ldots, n\}$ and consider the case where $l = l'$. Suppose that $|\eta_j| = \max_i |\eta_i|$. By proper normalization we can assume that $\eta_j = 1$. Let $y = \sum_{i \neq j} \eta_i a_i$. We can deduce the following properties for y:

$$\forall i \neq j :\ a_i \in T_{a_j}^{\perp} \Rightarrow y \in T_{a_j}^{\perp},$$
$$\|y\|^* = \max_{i \neq j} |\eta_i| \leq 1.$$

By the decomposability assumption $\sum_{i=1}^{k} \eta_i a_i = a_j + y \in \partial \|a_j\|$ hence $\left\| \sum_{i=1}^{k} \eta_i a_i \right\|^* \leq 1$. Hence $\left\| \sum_{i=1}^{k} \eta_i a_i \right\|^* = 1$. □

Remark. *Let $x = \sum_{i=1}^{K(x)} \gamma_i a_i$. Since $T_x^{\perp} = \bigcap_{i=1}^{K(x)} T_{a_i}^{\perp}$, a more general version of lemma 7 holds:*

Lemma 8. *Suppose $x \in \mathbb{R}^n$ and $y \in T_x^{\perp} - \{0\}$. If $z \in \mathcal{B}_{\|\cdot\|}$ is such that $\|y\|^* = \langle y, z \rangle$, then $z \in T_x^{\perp}$.*

We state and prove a dual version of Lemma 8, which will be used in the proof of Lemma 1 and Lemma 2.

Lemma 9. *Let $x \in \mathbb{R}^n$. If $y \in T_x^\perp$, then there exists $z \in T_x^\perp \cap \mathcal{B}_{\|\cdot\|^*}$ such that $\|y\| = \langle y, z\rangle$.*

Proof. If $y = 0$, then the lemma is trivially true. If $y \neq 0$, then:

$$\frac{y}{\|y\|} \in T_x^\perp \cap \{x \mid \|x\| = 1\} \Rightarrow \exists z \in T_x^\perp \text{ such that } \frac{y}{\|y\|} \in \underset{a \in T_x^\perp \cap \mathcal{B}_{\|\cdot\|}}{\operatorname{argmax}} \langle a, z\rangle.$$

Therefore, by Lemma 8, we get

$$\|z\|^* = \max_{a \in T_x^\perp \cap \mathcal{G}_{\|\cdot\|}} \langle a, z\rangle \leq \langle \frac{y}{\|y\|}, z\rangle \leq \|z\|^* \Rightarrow \langle \frac{y}{\|y\|}, z\rangle = \|z\|^* \Rightarrow \langle y, \frac{z}{\|z\|^*}\rangle = \|y\|$$

□ □

B.2 Proof of Theorem 2

First, we introduce a lemma.

Lemma 10. *Let $\{a_1, \ldots, a_k\}$ be an orthogonal subset of $\mathcal{G}_{\|\cdot\|}$ that satisfies II in Theorem 1. Let $y = \sum_{i=1}^k \beta_i a_i$, with $\beta_i \in \mathbb{R}$ for all i, then*

$$K(y) = |\{i \mid \beta_i \neq 0\}|.$$

Proof. Let $k' = |\{i \mid \beta_i \neq 0\}|$. Without loss of generality assume that $\beta_i \neq 0$ for $i \leq k'$ and $\beta_i = 0$ for $i > k'$. Let $\eta_i = \operatorname{sgn}(\beta_i)$ and $a_i' = \operatorname{sgn}(\beta_i) a_i$ for all $i \leq k'$. Since $a_1, \ldots, a_k$ satisfy condition II in the orthogonal representation theorem, so do $a_1', \ldots, a_{k'}'$.

Now we show that y and $a_1', \ldots, a_{k'}'$ satisfy condition I. By (2.11), $\left\|\sum_{i=1}^{k'} a_i'\right\|^* \leq 1$. Therefore,

$$\|y\| \geq \langle \sum_{i=1}^{k'} a_i', y\rangle = \sum_{i=1}^{k'} |\beta_i|, \quad \|y\| = \left\|\sum_{i=1}^{k'} \beta_i a_i'\right\| \leq \sum_{i=1}^{k'} |\beta_i| \quad \Rightarrow \quad \|y\| = \sum_{i=1}^{k'} |\beta_i|.$$

Therefore, by the orthogonal representation theorem, $e_y = \sum_{i=1}^{k'} a_i'$. Thus $K(y) = \|e_y\|_2^2 = k'$. □ □

For any $x \in \mathbb{R}^n - \{0\}$ define

$$l(x) = \min\{l \mid x = \sum_{i=1}^{l} \alpha_i b_i,\ b_1, \ldots, b_l \subseteq \mathcal{G}_{\|\cdot\|},\ \alpha_i \in \mathbb{R}\}.$$

Define $l(0) = 0$. Now the proof is a simple consequence of the following lemma:

Lemma 11. *For all $x \in \mathbb{R}^n$, $l(x) = K(x)$.*

Proof. $K(x) \geq l(x)$ by the definition of $l(x)$. We prove that $K(x) = l(x)$ by induction on $K(x)$. When $K(x) \in \{0, 1\}$, the statement is trivially true. Suppose the statement is true when $K(x) \in \{0, 1, 2, \ldots, k-1\}$. Consider the case where $K(x) = k$. By way of contradiction, suppose $l(x) < K(x)$. Let

$$x = \sum_{i=1}^{k} \gamma_i a_i, \tag{B.11}$$

where $\gamma_1, \ldots, \gamma_k$ and $a_1, \ldots, a_k$ are given by the orthogonal representation theorem. If $l(x) = 1$, then:

$$\sum_{i=1}^{k} \gamma_i a_i = \alpha_1 b_1,$$

for some $\alpha_1 \neq 0$ and $b_1 \in \mathcal{G}_{\|\cdot\|}$. Since $|\alpha_1| = \|\alpha_1 b_1\| = \|x\| = \sum_{i=1}^{k} \gamma_i$, either b_1 or $-b_1$ can be written as convex combination of $a_1, \ldots, a_k$ which contradicts the fact that $b_1 \in \mathcal{G}_{\|\cdot\|}$.

If $l(x) = l > 1$, we can write x as:

$$x = \sum_{i=1}^{l} \alpha_i b_i, \tag{B.12}$$

with $\{b_1, \ldots, b_l\} \subseteq \mathcal{G}_{\|\cdot\|}$. By turning b_i to $-b_i$ without loss of generality we assume that $\alpha_i > 0$ for all i. Let $u = 2\alpha_1 b_1$ and $v = 2\sum_{i=2}^{l} \alpha_i b_i$ and note that $x = (u + v)/2$. Let $C = \text{Cone}\{a_1, a_2, \ldots, a_k\}$. Let int$C$ and bd$C$ denote the interior and the boundary of C, respectively. Note that $u \notin$ intC because by Lemma 10, if $u \in$ intC, then $K(u) = k$; however, $l(u) = 1$. Now we consider two cases for v.

Case 1. If $v \in \text{intC}$, then we can write v as a conic combination of $a_1, a_2, \ldots, a_k$ with positive coefficients:

$$v = 2\sum_{i=2}^{l} \alpha_i b_i = \sum_{i=1}^{k} c_i a_i,$$

where $c_i > 0$ for all i.

Case 2. If $v \notin \text{int}C$. let $L = \{\theta u + (1-\theta)v \mid \theta \in [0,1]\}$. Since L intersects the interior of C at x and $\{u, v\} \notin \text{int}C$, there exists u', v' such that $L \cap \text{bd}C = \{u', v'\}$. Suppose v' is on the line segment between v and x (see Figure B.1). Let $L' = \{\theta u + (1-\theta)v' \mid \theta \in [0,1]\}$ and note that $x \in L'$. Since $v' \in \text{bd}C$, it can be written as conic combination of at most $k-1$ of $a_1, \ldots, a_k$. Without loss of generality assume that $v' = \sum_{i=2}^{k} \beta_i a_i$. For some $\theta \in (0,1)$:

$$x = \theta u + (1-\theta)v' = \alpha_1' b_1 + \sum_{i=2}^{k} \beta_i' a_i,$$

where $\alpha_1' = 2\theta\alpha_1$ and $\beta_i' = (1-\theta)\beta_i$. Using the representation in (B.11), we get:

$$\alpha_1' b_1 = \gamma_1 a_1 + \sum_{i=2}^{k} (\gamma_i - \beta_i') a_i.$$

We have $l(\alpha_1' b_1) = 1$, and by Lemma 10, $K(\alpha_1' b_1) = 1 + |\{i | \gamma_i \neq \beta_i', i = 2, \ldots, k\}|$. Therefore, $\gamma_i = \beta_i'$ for all $i = 2, \ldots, k$ and $b_1 = a_1$. Combining the previous fact with (B.11) and (B.12), we get:

$$x - \alpha_1 a_1 = (\gamma_1 - \alpha_1) a_1 + \sum_{i=2}^{k} \gamma_i a_i = \sum_{i=2}^{l} \alpha_i b_i. \tag{B.13}$$

If $\gamma_1 = \alpha_1$, by the induction hypothesis $k = l$, which is a contradiction. Now, suppose $\gamma_1 - \alpha_1 \neq 0$. In both cases we produced a point $y = v$ such that $K(y) = k$ and $l(y) \leq l - 1$. We can continues this procedure until we get a y such that $K(y) = k$ and $l(y) = 1$, which gives us the contradiction. $\square$

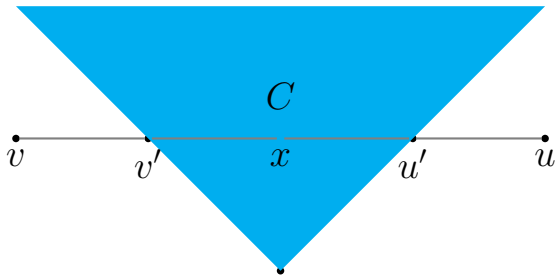


Figure B.1: Relative position of u' and v' on the line segment between u and v.

□

B.3 Proof of Proposition 2

In iteration $t+1$ when the backtrack procedure stops, the following inequality holds true:

$$\begin{aligned}\phi_\lambda(x^{(t+1)}) \leq m_{M_{t+1}}(x^{(t)}, x^{(t+1)}) &= \min_x f(x^{(t)}) + \langle \nabla f(x^{(t)}), x - x^{(t)} \rangle + \frac{M_{t+1}}{2}\left\|x - x^{(t)}\right\|_2^2 + \lambda\|x\| \\ &\leq \min_x \phi_\lambda(x) + \frac{M_{t+1}}{2}\left\|x - x^{(t)}\right\|_2^2. \end{aligned} \tag{B.14}$$

On the other hand, by (2.20), we have

$$\phi_\lambda(x^{(t+1)}) \leq m_{L_f}(x^{(t)}, x^{(t+1)}),$$

which ensures $M_{t+1} \leq \gamma_{\text{inc}} L_f$ since $m_L(x^{(t)}, x^{(t+1)})$ is non-decreasing in L. By (2.18), we have:

$$\phi_\lambda(x^{(t)}) \geq \phi_\lambda(x^*) + \frac{\mu_f}{2}\left\|x^{(t)} - x^*\right\|_2^2. \tag{B.15}$$

If we confine x to $\{\alpha x^* + (1-\alpha)x^{(t)} \,|\, 0 \leq \alpha \leq 1\}$, inequality (B.14) combined with (B.15)

results in

$$\begin{aligned}\phi_\lambda(x^{(t+1)}) &\leq \min_{\alpha\in[0,1]}\{\phi_\lambda(\alpha x^* + (1-\alpha)x^{(t)}) + \frac{\alpha^2 M_{t+1}}{2}\left\|x^{(t)} - x^*\right\|_2^2\} \\ &\leq \min_{\alpha\in[0,1]}\{\alpha\phi_\lambda(x^*) + (1-\alpha)\phi_\lambda(x^{(t)}) + \frac{\alpha^2 M_{t+1}}{2}\left\|x^{(t)} - x^*\right\|_2^2\} \\ &\leq \min_{\alpha\in[0,1]}\{\alpha\phi_\lambda(x^*) + (1-\alpha)\phi_\lambda(x^{(t)}) + \frac{\alpha^2\gamma_{\text{inc}}L_f}{\mu_f}(\phi_\lambda(x^{(t)}) - \phi_\lambda(x^*))\}.\end{aligned}$$

The RHS of the above inequality is minimized for $\alpha^* = \min\{1, \frac{\mu_f}{2\gamma_{\text{inc}}L_f}\}$. Therefore, we get

$$\phi_\lambda(x^{(t+1)}) - \phi_\lambda(x^*) \leq (1 - \alpha^* + \frac{\alpha^{*2}\gamma_{\text{inc}}L_f}{\mu_f})(\phi_\lambda(x^{(t)}) - \phi_\lambda(x^*)) \leq (1 - \frac{\mu_f}{4\gamma_{\text{inc}}L_f})(\phi_\lambda(x^{(t)}) - \phi_\lambda(x^*)).$$

To prove (2.24), we note that the backtrack stopping criteria ensures

$$\begin{aligned}\phi_\lambda(x^{(t+1)}) &\leq f(x^{(t)}) + \langle\nabla f(x^{(t)}), x^{(t+1)} - x^{(t)}\rangle + \frac{M_{t+1}}{2}\left\|x^{(t+1)} - x^{(t)}\right\|_2^2 + \lambda\left\|x^{(t+1)}\right\| \\ &\leq f(x^{(t)}) - \langle M_{t+1}(x^{(t+1)} - x^{(t)}) + \xi, x^{(t+1)} - x^{(t)}\rangle + \frac{M_{t+1}}{2}\left\|x^{(t+1)} - x^{(t)}\right\|_2^2 + \lambda\left\|x^{(t+1)}\right\| \\ &\leq f(x^{(t)}) - \frac{M_{t+1}}{2}\left\|x^{(t+1)} - x^{(t)}\right\|_2^2 + \langle\xi, x^{(t)} - x^{(t+1)}\rangle + \lambda\left\|x^{(t+1)}\right\| \\ &\leq \phi_\lambda(x^{(t)}) - \frac{M_{t+1}}{2}\left\|x^{(t+1)} - x^{(t)}\right\|_2^2. \qquad \text{(B.16)}\end{aligned}$$

The hypothesis (2.19) ensures $M_{t+1} \geq \mu_f$. Combining (2.17) and (B.16) and using the lower and the upper bounds on M_{t+1}, we get the desired result

$$\begin{aligned}\omega_\lambda(x^{(t+1)}) &\leq \left\|M_{t+1}(x^{(t)} - x^{(t+1)}) + \nabla f(x^{(t+1)}) - \nabla f(x^{(t)})\right\|^* \\ &\leq \theta(M_{t+1} + L'_f)\left\|x^{(t+1)} - x^{(t)}\right\|_2 \\ &\leq \theta(1 + \frac{L'_f}{M_{t+1}})\sqrt{2M_{t+1}(\phi_\lambda(x^{(t)}) - \phi_\lambda(x^{(t+1)}))} \\ &\leq \theta(1 + \frac{L'_f}{\mu_f})\sqrt{2\gamma_{\text{inc}}L_f(\phi_\lambda(x^{(t)}) - \phi_\lambda(x^*))}.\end{aligned}$$

B.4 Proof of Lemma 1

By the hypothesis there exists $\xi \in \partial\|x\|$ such that $\|A^*(Ax-b)+\lambda\xi\|^* \leq \delta\lambda$. Therefore, we can write

$$\begin{aligned}\delta\lambda\|x-x_0\| &\geq \|x-x_0\|\|A^*(Ax-b)+\lambda\xi\|^* \geq \langle (x-x_0), A^*(Ax-b)+\lambda\xi\rangle \\ &= \langle (x-x_0), A^*(A(x-x_0)) - A^*z + \lambda\xi\rangle \\ &= \|A(x-x_0)\|_2^2 - \langle x-x_0, A^*z\rangle + \lambda\langle x-x_0, \xi\rangle \\ &\geq \|A(x-x_0)\|_2^2 - \|x-x_0\|\|A^*z\|^* + \lambda(\|x\|-\|x_0\|). \end{aligned} \tag{B.17}$$

Now we lower-bound $\|x\|$:

$$\|x\| = \|x-x_0+x_0\| \geq \left\|P_{T_{x_0}{}^\perp}(x-x_0)+x_0\right\| - \left\|P_{T_{x_0}}(x-x_0)\right\|.$$

By Lemma 9, there exists $s \in T_{x_0}^\perp$ such that $\langle s, P_{T_{x_0}{}^\perp}(x-x_0)\rangle = \left\|P_{T_{x_0}{}^\perp}(x-x_0)\right\|$ and $\|s\|^* = 1$. Note that $e_{x_0}+s \in \partial\|x_0\|$ hence $\|e_{x_0}+s\|^* \leq 1$. Therefore, we get:

$$\left\|P_{T_{x_0}{}^\perp}(x-x_0)+x_0\right\| \geq \langle e_{x_0}+s, P_{T_{x_0}{}^\perp}(x-x_0)+x_0\rangle \geq \left\|P_{T_{x_0}{}^\perp}(x-x_0)\right\| + \|x_0\|,$$

$$\|x\| - \|x_0\| \geq \left\|P_{T_{x_0}{}^\perp}(x-x_0)\right\| - \left\|P_{T_{x_0}}(x-x_0)\right\|. \tag{B.18}$$

Combining (B.18) and (B.17), we get

$$\delta\lambda\|x-x_0\| \geq \lambda(\left\|P_{T_{x_0}{}^\perp}(x-x_0)\right\| - \left\|P_{T_{x_0}}(x-x_0)\right\|) - \|x-x_0\|\|A^*z\|^* + \|A(x-x_0)\|_2^2.$$

By applying triangle inequality to $\|x-x_0\|$, we obtain

$$(\lambda(1+\delta)+\|A^*z\|^*)\left\|P_{T_{x_0}}(x-x_0)\right\| \geq (\lambda(1-\delta)-\|A^*z\|^*)\left\|P_{T_{x_0}{}^\perp}(x-x_0)\right\| + \|A(x-x_0)\|_2^2. \tag{B.19}$$

That yields

$$\begin{aligned}\frac{\|x-x_0\|}{\|x-x_0\|_2} &\leq \frac{\left\|P_{T_{x_0}}(x-x_0)\right\| + \left\|P_{T_{x_0}{}^\perp}(x-x_0)\right\|}{\left\|P_{T_{x_0}}(x-x_0)\right\|_2} \\ &\leq (1+\gamma)\frac{\left\|P_{T_{x_0}}(x-x_0)\right\|}{\left\|P_{T_{x_0}}(x-x_0)\right\|_2} \leq (1+\gamma)\sqrt{ck_0}.\end{aligned}$$

Using the definition of the lower restricted isometry constant, we derive

$$\begin{aligned}\rho_-(A, c(1+\gamma)^2 k_0)\|x - x_0\|_2^2 \leq \|A(x-x_0)\|_2^2 &\leq^{(B.19)} ((1+\delta)\lambda + \|A^* z\|^*)\big\|P_{T_{x_0}}(x-x_0)\big\| \\ &\leq \sqrt{ck_0}((1+\delta)\lambda + \|A^* z\|^*)\big\|P_{T_{x_0}}(x-x_0)\big\|_2 \\ &\leq \sqrt{ck_0}((1+\delta)\lambda + \|A^* z\|^*)\|x - x_0\|_2,\end{aligned}$$

which yields the following bounds

$$\|x - x_0\|_2 \leq \frac{\sqrt{ck_0}((1+\delta)\lambda + \|A^* z\|^*)}{\rho_-(A, c(1+\gamma)^2 k_0)}, \tag{B.20}$$

$$\|x - x_0\| \leq \frac{ck_0(1+\gamma)((1+\delta)\lambda + \|A^* z\|^*)}{\rho_-(A, c(1+\gamma)^2 k_0)}. \tag{B.21}$$

By convexity of ϕ_λ,

$$\phi_\lambda(x) - \phi_\lambda(x_0) \leq \langle \lambda\xi + A^*(Ax - b), x - x_0\rangle \leq \frac{ck_0\delta\lambda(1+\gamma)((1+\delta)\lambda + \|A^* z\|^*)}{\rho_-(A, c(1+\gamma)^2 k_0)}.$$

B.5 Proof of Lemma 2

Let $\Delta = \frac{3ck_0\lambda(1+\gamma)}{2\rho_-(A, c(1+\gamma)^2 k_0)}$. We can write

$$\phi_\lambda(x) \leq \phi_\lambda(x_0) + \delta\lambda\Delta$$

$$\begin{aligned}\Rightarrow \quad \frac{1}{2}\|Ax - b\|_2^2 - \frac{1}{2}\|Ax_0 - b\|_2^2 &\leq \lambda(\|x_0\| - \|x\|) + \delta\lambda\Delta \\ &\leq \lambda\|x_0 - x\| + \delta\lambda\Delta \end{aligned} \tag{B.22}$$

If $\|x - x_0\| \leq \Delta$, half of the conclusion is immediate. To get the second half, we can expand the left hand side of (B.22) to get:

$$\begin{aligned}\frac{1}{2}\|A(x-x_0)\|_2^2 &\leq \lambda\|x - x_0\| + \langle x - x_0, A^* z\rangle + \delta\lambda\Delta \\ &\leq (\lambda + \|A^* z\|^*)\|x - x_0\| + \delta\lambda\Delta \\ &\leq (\frac{5}{4} + \delta)\lambda\Delta \leq \lambda\frac{3\Delta}{2}.\end{aligned}$$

Suppose $\|x - x_0\| > \Delta$, then from (B.22) we get:

$$
\begin{aligned}
f\lambda(\|x\| - \|x_0\|) &\leq \frac{1}{2}\|Ax_0 - b\|_2^2 - \frac{1}{2}\|Ax - b\|_2^2 + \delta\lambda\|x - x_0\| \\
&\leq -\frac{1}{2}\|A(x - x_0)\|_2^2 + \langle x - x_0, A^*z\rangle + \delta\lambda\|x - x_0\| \\
&\leq -\frac{1}{2}\|A(x - x_0)\|_2^2 + \|A^*z\|^*\|x - x_0\| + \delta\lambda\|x - x_0\|.
\end{aligned}
$$

By using (B.18) and triangle inequality we get:

$$
(\lambda(1+\delta) + \|A^*z\|^*)\left\|P_{T_{x_0}}(x - x_0)\right\| \geq (\lambda(1-\delta') - \|A^*z\|^*)\left\|P_{T_{x_0}{}^{\perp}}(x - x_0)\right\| + \frac{1}{2}\|A(x - x_0)\|_2^2.
$$

Using the same reasoning as in the proof of Lemma 1, we get the desired results.

B.6 Proof of Lemma 3

By first order optimality condition there exists $\xi \in \partial\|x^+\|$ such that:

$$
\begin{aligned}
\lambda\xi &= L(x - x^+) - \nabla f(x) \\
&= L(x - x^+) - A^*(Ax - b) \\
&= L(x - x^+) - A^*(A(x - x_0)) + A^*z
\end{aligned}
$$

Note that $\xi = e_{x^+} + v$ for some $v \in T_{x^+}^{\perp}$. By Lemma 9, there exists $v' \in T_{x^+}^{\perp} \cap \mathcal{B}_{\|\cdot\|^*}$ such that $\langle v', v\rangle = \|v\|$. Since $e_{x^+} + v' \in \partial\|x^+\|$, $\|e_{x^+} + v'\|^* \leq 1$. Therefore, we can write:

$$
\|\xi\| = \|e_{x^+} + v\| \geq \langle e_{x^+} + v', e_{x^+} + v\rangle = \|e_{x^+}\| + \|v\| \Rightarrow K(x^+) = \|e_{x^+}\| \leq \|\xi\|.
$$

Let $\xi = \sum_{i=1}^{l} \gamma_i a_i$, where $a_1, \ldots, a_l$ and $\gamma_1, \ldots, \gamma_l$ are given by the orthogonal representation theorem. Since $\gamma_i \leq 1$ for all i, $l \geq \|\xi\|$. If $\|\xi\| > \tilde{k}$, we can define $u = \sum_{i=1}^{\tilde{k}} a_i$, then

$$
\begin{aligned}
\tilde{k}\lambda \leq \langle u, \lambda\xi\rangle &= \langle u, L(x^+ - x)\rangle - \langle Au, A(x - x_0)\rangle + \langle u, A^*z\rangle \\
&\leq L\left\|x^+ - x\right\| + \sqrt{\rho_+(A, \tilde{k})\tilde{k}}\|A(x - x_0)\|_2 + \tilde{k}\|A^*z\|^* \\
\Rightarrow \frac{3\tilde{k}\lambda}{4} &\leq L\left\|x^+ - x\right\| + \sqrt{\rho_+(A, \tilde{k})\tilde{k}}\|A(x - x_0)\|_2. \qquad \text{(B.23)}
\end{aligned}
$$

Since $\phi_\lambda(x^+) \leq \phi_\lambda(x)$, by Lemma 2, we have:

$$\left\|x^+ - x\right\| \leq \left\|x^+ - x_0\right\| + \|x - x_0\| \leq \frac{9ck_0\lambda(1+\gamma)}{\rho_-(A, c(1+\gamma)^2k_0)},$$
$$\|A(x - x_0)\|_2^2 \leq \frac{9ck_0\lambda^2(1+\gamma)}{\rho_-(A, c(1+\gamma)^2k_0)}.$$

Define

$$\alpha = \gamma_{\text{inc}}\rho_+(A, 2\tilde{k})\frac{9ck_0(1+\gamma)}{\rho_-(A, c(1+\gamma)^2k_0)},$$
$$\beta^2 = \rho_+(A, \tilde{k})\frac{9ck_0(1+\gamma)}{\rho_-(A, c(1+\gamma)^2k_0)}.$$

We can rewrite (B.23) as:

$$\frac{3\tilde{k}}{4} - \alpha - \beta\sqrt{\tilde{k}} < 0 \Rightarrow \sqrt{\tilde{k}} < \frac{2}{3}(\beta + \sqrt{\beta^2 + 3\alpha}) \leq 2\sqrt{\alpha}.$$

But this contradicts Assumption 1, so $\|\xi\| \leq \tilde{k}$ hence $K(x^+) \leq \tilde{k}$.

Appendix C

PROOFS FROM CHAPTER 3

C.1 Proof of Lemma 4:

Using the definition of D_{sim}, we can write:

$$\begin{aligned} D_{\text{sim}} &= \sum_{t=1}^{m} \sigma_t(A_t^T \tilde{y}_t) - \psi^*(\tilde{y}_{m+1}) \\ &= \sum_{t=1}^{m} \langle A_t \tilde{x}_t, \tilde{y}_t \rangle - \psi^*(\tilde{y}_{m+1}) \\ &\leq \sum_{t=1}^{m} (\psi(\sum_{s=1}^{t} A_s \tilde{x}_s) - \psi(\sum_{s=1}^{t-1} A_s \tilde{x}_s)) - \psi^*(\tilde{y}_{m+1}) \\ &= \psi(\sum_{s=1}^{m} A_s \tilde{x}_s) - \psi(0) - \psi^*(\tilde{y}_{m+1}), \end{aligned}$$

where in the inequality follows from concavity of ψ, and the last line results from the sum telescoping. Similarly, we can bound D_{seq}:

$$\begin{aligned} D_{\text{seq}} = \sum_{t=1}^{m} \sigma_t(A_t^T \hat{y}_t) - \psi^*(\hat{y}_{m+1}) &= \sum_{t=1}^{m} \langle A_t \hat{x}_t, \hat{y}_t \rangle - \psi^*(\hat{y}_{m+1}) \\ &= \sum_{t=1}^{m} \langle A_t \hat{x}_t, -\hat{y}_{t+1} + \hat{y}_t \rangle + \sum_{t=1}^{m} \langle A_t \hat{x}_t, \hat{y}_{t+1} \rangle - \psi^*(\hat{y}_m) \\ &\leq \sum_{t=1}^{m} \langle A_t \hat{x}_t, -\hat{y}_{t+1} + \hat{y}_t \rangle + \sum_{t=1}^{m} (\psi(\sum_{s=1}^{t} A_s \hat{x}_s) - \psi(\sum_{s=1}^{t-1} A_s \hat{x}_s)) - \psi^*(\hat{y}_{m+1}) \\ &= \sum_{t=1}^{m} \langle A_t \hat{x}_t, -\hat{y}_{t+1} + \hat{y}_t \rangle + \psi(\sum_{s=1}^{m} A_s \hat{x}_s) - \psi(0) - \psi^*(\hat{y}_{m+1}). \end{aligned} \tag{C.1}$$

When ψ is differentiable with Lipschitz gradient, we can use the following inequality that is equivalent to Lipschitz continuity of the gradient:

$$\psi(u') \geq \psi(u) + \langle \nabla\psi(u), u' - u\rangle - \frac{1}{2\mu}\|u - u'\|^2 \quad u, u' \in K$$

(see, for example, [NN04, section 2.1.1]) to get

$$\begin{aligned} D_{\text{seq}} = \sum_{t=1}^{m} \sigma_t(A_t^T \hat{y}_t) - \psi^*(\hat{y}_{m+1}) &= \sum_{t=1}^{m} \langle A_t\hat{x}_t, \hat{y}_t\rangle - \psi^*(\hat{y}_{m+1}) \qquad \text{(C.2)} \\ &\leq \sum_{t=1}^{m} \frac{1}{2\mu}\|A_t\hat{x}_t\|^2 + \sum_{t=1}^{m}(\psi(\sum_{s=1}^{t} A_s\hat{x}_s) - \psi(\sum_{s=1}^{t-1} A_s\hat{x}_s)) - \psi^*(\hat{y}_{m+1}) \\ &= \sum_{t=1}^{m} \frac{1}{2\mu}\|A_t\hat{x}_t\|^2 + \psi(\sum_{s=1}^{m} A_s\hat{x}_s) - \psi(0) - \psi^*(\hat{y}_{m+1}). \end{aligned}$$

□

C.2 Proof of Lemma 5:

Let (y, β) be a feasible solution for problem (3.49). Note that $y \geq 0$ since $\mathbf{dom}\ \psi^* \subset \mathbb{R}_+$ by the fact that ψ is non-decreasing. Let $\bar{y}(t) = \inf_{s\leq t} y(s)$. Note that $\bar{y}$ is continuous. Define

$$\beta(t) = \frac{\int_{s=0}^{t} y(s)\ ds - \psi^*(y(t))}{\psi(t)}, \qquad \bar{\beta}(t) = \frac{\int_{s=0}^{t} \bar{y}(s)\ ds - \psi^*(\bar{y}(t))}{\psi(t)},$$

with the definition modified with the right limit at $t = 0$. For any t such that $\bar{y}(t) = y(t)$, we have:

$$\beta(t) = \frac{\int_{s=0}^{t} y(s)\ ds - \psi^*(y(t))}{\psi(t)} \geq \frac{\int_{s=0}^{t} \bar{y}(s)\ ds - \psi^*(\bar{y}(t))}{\psi(t)} = \bar{\beta}(t).$$

Now, we consider the set $\{t \mid \bar{y}(t) \neq y(t)\}$. By the definition of $\bar{y}$, we have $\bar{y}(0) = y(0)$. Since both functions are continuous, the set $\{t \mid \bar{y}(t) \neq y(t)\}$ is an open subset of $\mathbb{R}$ and hence can be written as a countable union of disjoint open intervals. Specifically, we can define the end points of the intervals as:

$$\begin{aligned} &a_0 = b_0 = 0, \\ &a_i = \inf\{t > b_{i-1} \mid y(t) > \bar{y}(t)\}, \quad \forall i \in \{1, 2, \ldots\} \\ &b_i = \inf\{t > a_i \mid y(t) = \bar{y}(t)\}, \quad \forall i \in \{1, 2, \ldots\}. \end{aligned}$$

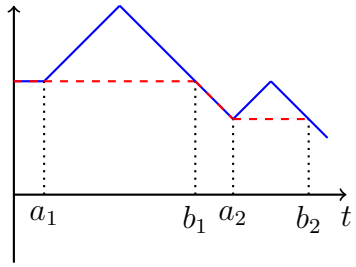


Figure C.1: An example of $y(t)$ (solid blue) and $\bar{y}(t)$ (dashed red).

then $\{t \mid \bar{y}(t) \neq y(t)\} = \bigcup_{i\in\{1,2,\ldots\}}(a_i,\ b_i)$.(See Figure C.1)

For any $i \in \{1, 2, \ldots\}$, we show that $\beta(t) \geq \bar{\beta}(t)$ on $(a_i,\ b_i)$. If $a_i = \infty$, then $b_i = \infty$, so we assume that $a_i < \infty$. By the definition of a_i and b_i, $\bar{y}(t)$ is constant on $(a_i,\ b_i)$. Also, we have $y(a_i) = \bar{y}(a_i)$. Similarly, we have $y(b_i) = \bar{y}(b_i)$ whenever $b_i < \infty$.

Since $\bar{y}(t) \leq y(t)$ for all t and $y(a_i) = \bar{y}(a_i)$, we have

$$\beta(a_i) \geq \bar{\beta}(a_i). \tag{C.3}$$

If $b_i < \infty$, similarly by the fact that $y(b_i) = \bar{y}(a_i) = \bar{y}(b_i)$, we have

$$\beta(b_i) \geq \bar{\beta}(b_i). \tag{C.4}$$

Now we consider the case where $b_i = \infty$. In this case we have $\bar{y}(t) = \bar{y}(a_i)$ on $(a_i,\ \infty)$. We consider two cases based on the asymptotic behavior of ψ. If $\lim_{t\to\infty} \psi(t) = +\infty$ (ψ is unbounded), then we have

$$\limsup_{t\to\infty} \beta(t) = \limsup_{t\to\infty} \frac{\int_{s=0}^{t} y(s)\ ds}{\psi(t)} \geq \limsup_{t\to\infty} \frac{\int_{s=0}^{t} \bar{y}(a_i)\ ds}{\psi(t)} = \lim_{t\to\infty} \bar{\beta}(t). \tag{C.5}$$

Here we used the fact that $-\psi^*(y(t))$ is bounded. This follows from the fact ψ^* is monotone thus:

$$-\psi^*(y(t)) \leq -\psi^*(\bar{y}(a_i)),$$

and $-\psi^*(\bar{y}(a_i)) < \infty$ because if $-\psi^*(\bar{y}(a_i)) = \infty$, then $\beta(a_i) \geq \bar{\beta}(a_i) = \infty$ which contradicts the feasibility of (y, β).

Now consider the case when $\lim_{t\to\infty} \psi(t) = M$ for some positive constant M. In this case, $-\psi^* \leq M$. We claim that $y(a_i) = 0$ and $\liminf_{t\to\infty} y(t) = 0$. Suppose $\liminf_{t\to\infty} y(t) > 0$, then $\limsup_{t\to\infty} \beta(t) = \infty$ since the numerator in the definition of β tends to infinity while the denominator is bounded. But this contradicts feasibility of (y, β). On the other hand, by the definition of a_i and b_i we should have $y(a_i) = \bar{y}(a_i) \leq \liminf_{t\to\infty} y_t$. Combining this with the fact that $\bar{y}(a_i) \in \mathbf{dom}\ \psi^* \subset \mathbb{R}_+$, we conclude that $y(a_i) = 0$. Using that $y(a_i) = 0$ and $\liminf_{t\to\infty} y(t) = 0$, we get:

$$\begin{aligned}
\limsup_{t\to\infty} \beta(t) &= \limsup_{t\to\infty} \frac{\int_{s=0}^{t} y(s)\ ds - \psi^*(y(t))}{\psi(t)} \\
&\geq \lim_{t\to\infty} \frac{\int_{s=0}^{t} y(s)\ ds - \psi^*(0)}{M} \\
&\geq \frac{\int_{s=0}^{a_i} \bar{y}(s)\ ds - \psi^*(0)}{M} = \lim_{t\to\infty} \bar{\beta}(t),
\end{aligned} \tag{C.6}$$

where in the last inequality we used the fact that $\bar{y}(t) = 0$ for $t \geq a_i$.

Let ψ' be the right derivative of ψ. Since ψ is concave, ψ' is non-increasing. Therefore, the interval (a_i, b_i) can be written as $(a_i, t'] \cup [t', b_i)$ such that $\psi'(t) \geq \bar{y}(a_i)$ on $(a_i, t']$ and $\psi'(t) \leq \bar{y}(a_i)$ on $[t', b_i)$. Since $\psi'(t) \geq \bar{y}(a_i)$ on $(a_i, t']$ we have:

$$\begin{aligned}
\int_{a_i}^{t} \bar{y}(t)\ dt &= \int_{a_i}^{t} \bar{y}(a_i)\ dt \\
&\leq \int_{a_i}^{t} \psi'(t)\ dt = \psi(t) - \psi(a_i),
\end{aligned}$$

for all $t \in (a_i, t']$. This yields:

$$\begin{aligned}
\bar{\beta}(a_i) &= \frac{\int_{s=0}^{a_i} \bar{y}(s)\ ds - \psi^*(\bar{y}(a_i))}{\psi(t)} \\
&\geq \frac{\int_{s=0}^{a_i} \bar{y}(s)\ ds + \int_{a_i}^{t} \bar{y}(t)\ dt - \psi^*(\bar{y}(a_i))}{\psi(a_i) + \psi(t) - \psi(a_i)} = \bar{\beta}(t).
\end{aligned}$$

for all $t \in (a_i, t']$. Here we used the fact that if $c_1 \geq c_2 > 0$ and $d_2 \geq d_1 \geq 0$, then

$$\frac{c_1}{c_2} \geq \frac{c_1 + d_1}{c_2 + d_2}.$$

Similarly, we have $\bar{\beta}(b_i) \geq \bar{\beta}(t)$ for any $t \in [t', b_i)$. Combining this with (C.3),(C.4),(C.5), and (C.6), we get:

$$\begin{aligned}\sup_{a_i \leq t \leq b_i} \bar{\beta}(t) &= \max(\bar{\beta}(a_i), \bar{\beta}(b_i)) \\ &\leq \max(\beta(a_i), \beta(b_i)) \leq \sup_{a_i \leq t \leq b_i} \beta(t).\end{aligned}$$

We conclude that $\bar{\beta}(t) \leq \beta(t)$ for all $t \geq 0$ hence $(\bar{y}, \beta)$ is a feasible solution for the problem.

C.3 Proof of Theorem 8:

Let (y, β) be a feasible solution for problem (3.49). By Lemma 5, we can assume that y is non-increasing. First, note that $y(t) \geq 0$ since $\mathbf{dom}\ \psi^* = [0,\ \infty)$. Define $\bar{y}(t) = y(t)$ for $t \leq t'$ and $\bar{y}(t) = 0$ for $t > t'$. We show that $(\bar{y}, \beta)$ is also a feasible solution for (3.49) modulo the continuity condition. Define

$$\beta(t) = \frac{\int_{s=0}^{t} y(s)\ ds - \psi^*(y(t))}{\psi(t)}, \qquad \bar{\beta}(t) = \frac{\int_{s=0}^{t} \bar{y}(s)\ ds - \psi^*(\bar{y}(t))}{\psi(t)}.$$

By the definition of $\bar{y}$, we have:

$$\int_0^t y(s)\ ds \geq \int_0^t \bar{y}(s)\ ds, \tag{C.7}$$

for all t and $\beta(t) = \bar{\beta}(t)$ for $t \in [0,\ t']$. Since $y(t)$ is non-increasing and $y(t) \geq 0$, $\lim_{t\to\infty} y(t)$ exists. We claim that $\lim_{t\to\infty} y(t) = 0$. To see this note that if $\lim_{t\to\infty} y(s) > 0$, then

$$\lim_{t\to\infty} \int_{s=0}^{t} y(s)\ ds = \infty,$$

which contradicts the fact that $\beta(t) \leq \beta$ for all t. For all $t \geq t'$, now we have:

$$\begin{aligned}\sup_{t \geq t'} \beta(t) \geq \lim_{t\to\infty} \beta(t) &= \frac{\lim_{t\to\infty} \int_{s=0}^{t} y(s)\ ds - \psi^*(0)}{\psi(t')} \\ &\geq \bar{\beta}(t'),\end{aligned}$$

where the equality follows from the fact that $\lim_{t\to\infty} y(t) = 0$, and in the last inequality, we used (C.7). Since $\bar{y}(t) = 0$ for $t > t'$, $\beta(t)$ is constant on $[t' \ \infty)$. Therefore, $\sup_{t\geq t'} \bar{\beta}(t) = \bar{\beta}(t')$. Combining this with the previous inequality we get:

$$\sup_{t\geq t'} \beta(t) \geq \sup_{t\geq t'} \bar{\beta}(t').$$

Therefore, we conclude that $\bar{\beta}(t) \leq \beta$ for all t. Thus $(\bar{y}, \beta)$ is also a feasible solution for (3.49) modulo the continuity condition. Note that $\bar{y}(t)$ may not be continuous at t'. However, we can find a sequence of continuous functions $z^{(j)}$ that converge pointwise to y and $z^{(i)}(t) = 0$ for all i and $t \geq t'$. To do so we consider a sequence of real number $\epsilon_i \to 0$. We define $z^{(i)}(t) = \bar{y}(t)$ for $t \in [0, \ t' - \epsilon_i) \cup [t', \ \infty)$. On $[t' - \epsilon_i, \ t']$ we define $z^{(i)}(t)$ to be a linear function that take values $y(t' - \epsilon)$ and 0 on the endpoints. Define

$$\beta_{z^{(i)}} = \sup_{t>0} \frac{\int_{s=0}^{t} z_s^{(i)} \, ds - \psi^*(z_t^{(i)})}{\psi(t)}.$$

By upper semi-continuity of ψ^*, $\beta_{z^{(i)}}$ converges to $\bar{\beta}$.

Let β^* be the optimal solution for problem (3.49). By the definition, there exits a feasible sequence $(y^{(j)}, \beta^{(j)})$ such that $\beta^{(j)}$ converges to β^*. Let $\bar{y}(t)^{(j)} = y(t)^{(j)}$ for $t \leq t'$ and $\bar{y}(t)^{(j)} = 0$ for $t > t'$. Note that $\bar{y}(t)^{(j)}$ may not be continuous at t'. However, we can find a sequence of continuous functions $(z^{(ji)}, \ \beta_{z^{(ji)}})$ as in above. Now $\beta_{z^{(jj)}}$ converges to β^*.

□

C.3.1 Distance from l_p norm ball

In this section we prove that the function:

$$G(u) = -d_1(u, \mathcal{B}_p)$$

satisfies Assumption 2 and find a lower bound on $\bar{\alpha}_\psi$ when ψ is given by 3.17 with $G(u) = -ld_1(u, \mathcal{B}_p)$.

For any $u \in \mathbb{R}^n_+$, there exists $\bar{u} \in \mathcal{B}_p$ such that $d_1(u, \mathcal{B}_p) = \|u - \bar{u}\|_1$. the subdifferential of distance function is[1]:

$$\partial d_1(u, \mathcal{B}_p) = \partial \|u - \bar{u}\|_1 \cap N_{\mathcal{B}_p}(\bar{u}),$$

where $N_{\mathcal{B}_p}(u) = \{\xi \mid \langle \xi, v - u \rangle \geq 0, \ \forall v \in \mathcal{B}_p\}$ is the normal cone of $\mathcal{B}_p$ at u. In fact $d_1(u, \mathcal{B}_p) = \|u - \bar{u}\|_1$ if and only if $\partial \|u - \bar{u}\|_1 \cap N_{\mathcal{B}_p}(\bar{u}) \neq \varnothing$. When $u \in \text{int}\mathcal{B}_p$, $\bar{u} = u$ and $\partial d_1(u, \mathcal{B}_p) = \{0\}$. In order to find $\partial d_1(u, \mathcal{B}_p)$ when $u \notin \text{int}\mathcal{B}_p$, we first find $\bar{u}$ in this case. For any $r \geq 0$, define $u \wedge r \in \mathbb{R}^n_+$ to be:

$$(u \wedge r)_i = \min(u_i, r) \quad \forall i.$$

Note that $\|u \wedge 0\|_p = 0$ and $\|u \wedge (\max_i u_i)\|_p = \|u\|_p \geq 1$. Since $\|u \wedge r\|_p$ is a continuous function of r, by the intermediate value theorem, there exists $r_u \in (0, \max_i u_i]$ such that $\|u \wedge r_u\|_p = 1$. Now $\bar{u} = u \wedge r$. To see this note that:

$$\partial \|u - \bar{u}\|_1 \cap N_{\mathcal{B}_p}(\bar{u}) = \left\{ \frac{1}{r_u^{p-1}} (u \wedge r_u)^{\circ(p-1)} \right\} \quad \text{for} \quad r_u < \max_i u_i; \tag{C.8}$$

$$\partial \|u - \bar{u}\|_1 \cap N_{\mathcal{B}_p}(\bar{u}) = \left\{ \frac{z}{r_u^{p-1}} (u \wedge r_u)^{\circ(p-1)} \mid 0 \leq z \leq 1 \right\} \quad \text{for} \quad r_u = \max_i u_i; \tag{C.9}$$

where $^{\circ(p-1)}$ denotes element-wise exponentiation. Now if $u' \leq u$, then $r_u \leq r'_u$ since $\|u \wedge r\|_p \geq \|u' \wedge r\|_p$ for all r. Thus by (C.8) and (C.9), there exists $y \in \partial d_1(u, \mathcal{B}_p)$ such that $y \geq \partial d_1(u', \mathcal{B}_p)$.

Now we can find $\bar{\alpha}_\psi$ when ψ is given by 3.17 with $G(u) = -l d_1(u, \mathcal{B}_p)$. Note that by (C.8) when $y \in \partial\psi(u)$ with $u \notin \mathcal{B}_p$, then $\min_i y_i = -l$. Now by the definition of l in (3.18) and the explanation that followed it, if $(\sum_{t=1}^m c_t x_t, \sum_{t=1}^m B_t x_t) \in Q_\psi$, then we must have $\sum_{t=1}^m B_t x_t \in \mathcal{B}_p$. Suppose $(\sum_{t=1}^m c_t x_t, \sum_{t=1}^m B_t x_t) \in Q_\psi$ and let $u = \sum_{t=1}^m B_t x_t$ and $v = \sum_{t=1}^m c_t x_t$. If $u \in \text{int}\mathcal{B}_p$, then G is differentiable at u and $\nabla G(u) = 0$ which yields $\alpha_\psi(v, u) = 0$. Now suppose $u \in \text{bd}\mathcal{B}_p$. In this case we have:

$$\alpha_\psi(v, u) = \min_{y \in \partial G(u)} \frac{G^*(y)}{\psi\left(\sum_{t=1}^m A_t x_t\right)} = \min_{y \in \partial G(u)} \frac{-\|y\|_q}{\sum_{t=1}^n c_t \bar{x}_t} \geq \min_{y \in \partial G(u)} \frac{-\|y\|_q}{\theta \mathbf{1}^T u}.$$

[1]For convex function we use ∂ to denote subdifferential.

Recall that $\theta = \min_t \min_{x \in F_t} \frac{c_t^T x}{\mathbf{1}^T B_t x}$. By (C.9), we have:

$$\min_{y \in \partial G(u)} -\|y\|_q = -\frac{l}{(\max_i u_i)^{p-1}}.$$

Therefore,

$$\alpha_\psi(v,u) \geq -\frac{l}{\theta} \frac{1}{(\mathbf{1}^T u)(\max_i u_i)^{p-1}}.$$

As u varies on the bd$\mathcal{B}_p$, the right hand side is lower bounded by $\frac{-l}{\theta}$. This yields $\alpha_\psi \geq -\frac{l}{\theta}$.

C.3.2 Derivation of lower bounds on $\bar{\alpha}_{\psi,\psi\Box\phi}$

We first derive a general inequality which will be specialized to different examples for bounding $\alpha_{\psi,\psi\Box\phi}$. Let $K = K_1 \times K_2$, with K_1 and K_2 two proper cones. Suppose $\psi(v,u) = H(v) + G(u)$, where $H : K_1 \mapsto \mathbb{R}$ is a non-decreasing, and $G : K_2 \mapsto \mathbb{R}$ is non-increasing and l Lipschitz continuous. We assume $H(v) \geq \theta u$ for all $(v,u) \in \sum_{t=1}^m A_t F_t$. Note that $\psi^*(z,y) = H^*(z) + G^*(y)$. We set

$$\phi^*(y) = \sum_{i=1}^m \frac{1}{\gamma} \left(\left(y_i - \frac{\theta}{(e-1)} \right) \log \left(1 - \frac{(e-1)}{\theta} y_i \right) - (1+\gamma) y_i \right),$$

where $\gamma = \log(1 + \frac{l(e-1)}{\theta})$. We let $\psi\Box\phi(v,u) = H(v) + G\Box\phi(u)$. Let $(v,u) \in Q_{\psi\Box\phi}$. Since $\psi\Box\phi(0) = 0$, by the definition of $Q_{\psi\Box\phi}$,

$$H(v) + G\Box\phi(u) \geq \psi\Box\phi(0) = 0. \tag{C.10}$$

Let $(z,y) \in \partial\psi\Box\phi(u)$, then we have:

$$\begin{aligned} u_i &= \nabla_i \phi^*(y) + \tilde{\nabla}_i G^*(y) \\ &= \frac{1}{\gamma} \log(1 + \frac{(e-1)y_i}{\theta}) - 1 + \tilde{\nabla}_i G^*(y), \end{aligned} \tag{C.11}$$

for some $\tilde{\nabla} G^*(y) \in \partial G^*(y)$. Using the previous identity, we can derive the following upper bound for $G\Box\phi(u)$:

$$
\begin{aligned}
G\Box\phi(u) &= \langle y, u\rangle - G^*(y) - \phi^*(y) \\
&= \langle y, \tilde{\nabla} G^*(y)\rangle - G^*(y) + \langle y, \nabla\phi^*(y)\rangle - \phi^*(y) \\
&= G(\tilde{\nabla} G^*(y)) + \phi(\nabla\phi^*(y)) \\
&= G(\tilde{\nabla} G^*(y)) + \frac{\theta}{(e-1)}\sum_{i=1}^{m}(u_i - \tilde{\nabla}_i G^*(y) + 1) + \frac{1}{\gamma}\mathbf{1}^T y \\
&\leq G(\tilde{\nabla} G^*(y)) + \frac{1}{(e-1)}H(v) + \frac{\theta}{(e-1)}\sum_{i=1}^{m}(1 - \tilde{\nabla}_i G^*(y)) + \frac{1}{\gamma}\mathbf{1}^T y. \qquad \text{(C.12)}
\end{aligned}
$$

Now we specialize the bound to the case where $G : \mathbf{R}_+^m \mapsto \mathbf{R}$ and $G(u) = -l\sum_{i=1}^{m}(u_i - 1)_+$. We assume $u \leq \mathbf{1}$ for all $(v, u) \in Q_{\psi\Box\phi}$. In that case, (C.11) is satisfied with $\tilde{\nabla}_i G^*(y) = \mathbf{1}$. Thus from (C.12) simplifies to:

$$G\Box\phi(u) \leq \frac{1}{(e-1)}H(v) + \frac{1}{\gamma}\mathbf{1}^T y. \qquad \text{(C.13)}$$

Combining this with (C.10), we get:

$$H(v) + G(u) = H(v) \geq \frac{(1/e-1)}{\gamma}\mathbf{1}^T y. \qquad \text{(C.14)}$$

In the view of definition of $\alpha_{\psi,\psi\Box\phi}$, by using (C.13) and the fact that $G^*(y) = \mathbf{1}^T y$, we derive the following inequality:

$$H(v) + G\Box\phi(u) - G^*(y) \leq (1 + \frac{1}{e-1})H(v) + (\frac{1}{\gamma} - 1)\mathbf{1}^T y. \qquad \text{(C.15)}$$

Combining (C.14) and (C.15), we get the following lower bound on $\alpha_{\psi,\psi\Box\phi}$:

$$\bar{\alpha}_{\psi,\psi\Box\phi} \geq 1 - (1 + \frac{1}{e-1})\gamma + \bar{\alpha}_H. \qquad \text{(C.16)}$$

C.4 Online LP:

appendix-lp In this problem,

$$\psi\left(\sum_{t=1}^{m} A_t x_t\right) = \sum_{t=1}^{m} c_t^T x_t + G\left(\sum_{t=1}^{m} B_t x_t\right), \qquad \text{(C.17)}$$

with $G(u) = -l\sum_{i=1}^{n}(u_i - 1)_+$. In this problem $H(v) = v$ is the identity function. Recall that l is defined such that:

$$l \geq \max\left\{ \frac{c_t^T x}{(B_t x)_i} \mid x \in F_t, (B_t x)_i > 0, i \in [m] \right\}.$$

Let $(v, u) = (\sum_{t=1}^{m} c_t^T x_t, \sum_{t=1}^{m} B_t x_t) \in Q_{\psi\Box\phi}$. By the definition of l and $Q_{\psi\Box\phi}$ we have $u \leq 1$. On the other hand, by the definition of θ, we have $H(v) \geq \theta u$. Since H is a linear function, $\alpha_H = 0$. Thus (C.16) yields

$$\bar{\alpha}_{\psi,\psi\Box\phi} \geq 1 - (1 + \frac{1}{e-1})\gamma. \tag{C.18}$$

C.5 Proof of Lemma 6

Proof. To apply Clarke's exact penalty principle, let $C = \{x \in \mathbf{R}^m : \sum_{t=1}^{m} x_t \leq b\}$ and note that the distance from x to C in infinity norm is exactly $(\sum_{t=1}^{m} x_t - b)_+$. Then it is enough to check that the function $x \mapsto H(\sum_{t=1}^{m} A_t x_t)$ is l-Lipschitz with respect to infinity norm (see [Bur91, Theorem 5.5]). This holds because

$$\frac{\partial}{\partial x_t} H\left(\sum_{t=1}^{m} A_t x_t\right) = \langle A_t, \nabla H\left(\sum_{t=1}^{m} A_t x_t\right)\rangle \leq h'\left(\lambda_{\min}\left(\sum_{t=1}^{m} A_t x_t\right)\right) \mathbf{tr} A_t \leq h'(0)\, \mathbf{tr} A_t \leq l.$$

Here we have used the fact that h' is monotonically decreasing (because h is concave), so the largest eigenvalue of $\nabla H(U)$ is $h'(\lambda_{\min}(U))$, and $h'(u) \geq h'(0)$ for all $u \geq 0$. □

C.5.1 Monotonicity of left-hand side of (3.48)

Lemma 12. *If $h'(v) \geq 0$ for all $v \in [0, b]$ then the function $F : (0, \infty) \to \mathbf{R}$ defined by*

$$F(\gamma) = \frac{\gamma}{b} \int_0^b \exp\left(\gamma\left(1 - \frac{v}{b}\right)\right) h'(v)\, dv$$

is monotonically increasing.

Proof. Computing the derivative of F we obtain

$$F'(\gamma) = \frac{1}{b} \int_0^b \exp\left(\gamma\left(1 - \frac{v}{b}\right)\right) h'(v)\, dv + \frac{\gamma}{b} \int_0^b \left(1 - \frac{v}{b}\right) \exp\left(\gamma\left(1 - \frac{v}{b}\right)\right) h'(v)\, dv.$$

Since $h'(v) \geq 0$ for all $v \in [0, b]$ and $1 - v/b \geq 0$ for all $v \in [0, b]$ it follows that $F'(\gamma) \geq 0$ as required. □

C.6 Proof of Theorem 10

Proof. Let $U = \sum_{t=1}^{m} A_t \hat{x}_t$, $u = \sum_{t=1}^{m} \hat{x}_t$, $Y = \nabla H_S(U)$, and $z = G_S'(u)$.

First we show that $u \leq b$. If $\hat{z}_t \leq -l$ for some t, then since $\langle \hat{Y}_t, A_t \rangle \leq h'(0)\mathbf{tr}(A_t) < l$ (by the same argument as in the proof of Lemma 6) it follows that $\langle \hat{Y}_t, A_t \rangle + \hat{z}_t < 0$ which results in $\hat{x}_{t+1} = 0$. Given that $G_S{}'(u) = -l$ when $u \geq b - \rho_1$, we conclude $u \leq b$ and therefore, $G(u) = -l(u-b)_+ = 0$.

Note that since $\hat{z}_t \leq \hat{z}_{t-1}$ and $\hat{Y}_t \preceq \hat{Y}_{t-1}$, we can bound the extra terms in (**??**) as:

$$\sum_{t=1}^{m} \langle A_t \hat{x}_t, \hat{Y}_{t-1} - \hat{Y}_t \rangle \leq \rho_1 \rho_2 \sum_{t=1}^{m} \mathbf{tr}\left(\hat{Y}_{t-1} - \hat{Y}_t\right) \leq \rho_1 \rho_2 \left(\mathbf{tr}\left(\hat{Y}_0\right) - \mathbf{tr}\left(\hat{Y}_S\right)\right) \quad \text{and} \tag{C.19}$$

$$\sum_{t=1}^{m} \hat{x}_t (\hat{z}_{t-1} - \hat{z}_t) \leq \rho_1 \sum_{t=1}^{m} (\hat{z}_{t-1} - \hat{z}_t) = \rho_1 (\hat{z}_0 - \hat{z}_m) = -\rho_1 \hat{z}_m \tag{C.20}$$

Also by the primal allocation rule we have:

$$\hat{x}_t \left(\hat{z}_{t-1} + \langle A_t, \hat{Y}_{t-1} \rangle\right) \geq 0.$$

Combining this by the concavity of H_S and G_S, we get:

$$H_S\left(\sum_{s=1}^{t} A_t \hat{x}_s\right) + G_S\left(\sum_{s=1}^{t} \hat{x}_s\right) - H_S\left(\sum_{s=1}^{t-1} A_t \hat{x}_s\right) - G_S\left(\sum_{s=1}^{t-1} \hat{x}_s\right) + \hat{x}_t \left(\hat{z}_{t-1} - \hat{z}_t + \langle \hat{Y}_{t-1}, A_t \rangle - \langle \hat{Y}_t, A_t \rangle\right) \geq 0.$$

By taking the sum over t and telescoping the sum we get:

$$H_S\left(\sum_{s=1}^{m} A_t(\hat{x}_s)\right) + G_S\left(\sum_{s=1}^{m} \hat{x}_s\right) + \sum_{t=1}^{m} \hat{x}_t \left(\hat{z}_{t-1} - \hat{z}_t + \langle A_t, \hat{Y}_{t-1} - \hat{Y}_t \rangle\right) \geq 0. \tag{C.21}$$

The proof follows the same step as the proof of Theorem 9 and uses (4) and the above

inequalities.

$$
\begin{aligned}
&P_{\text{seq}} - D_{\text{seq}} \\
&= H(U) + G(u) - D_{\text{seq}} \\
&\text{By Lemma (4)} \\
&\geq -H_S(U) - G_S(u) + H^*(Y) + G^*(z) + \sum_{t=1}^{m} \hat{x}_t \left(\langle A_t, \hat{Y}_t - \hat{Y}_{t-1} \rangle + \hat{z}_t - \hat{z}_{t-1} \right) + H(U) + G(u) \\
&\text{By (3.51)} \\
&\geq -H_S(U) + (\gamma - 1) G_S(u) + (1 - \gamma) H(U) + H^*(Y) - \gamma \rho_1 G'_S(u) \\
&\quad + \sum_{t=1}^{m} \hat{x}_t \left(\langle A_t, \hat{Y}_t - \hat{Y}_{t-1} \rangle + \hat{z}_t - \hat{z}_{t-1} \right) \\
&\text{By (C.21)} \\
&\geq -H_S(U) + (1 - \gamma) H_S(U) + (1 - \gamma) H(U) + H^*(Y) - \gamma \rho_1 G'_S(u) \\
&\quad + \gamma \sum_{t=1}^{m} \hat{x}_t \left(\langle A_t, \hat{Y}_t - \hat{Y}_{t-1} \rangle + \hat{z}_t - \hat{z}_{t-1} \right) \\
&\text{By (C.19) \& (C.20)} \\
&\geq -H_S(U) + (1 - \gamma) H_S(U) + (1 - \gamma) H(U) + H^*(Y) - \rho_1 \rho_2 \left(\mathbf{tr}\left(\hat{Y}_0 \right) - \mathbf{tr}\left(\hat{Y}_S \right) \right) \\
&\text{By (3.50)} \\
&\geq (1 - \gamma - \beta) H(U) \\
&= (1 - \gamma - \beta) P_{\text{seq}}
\end{aligned}
$$

□