Randomized Subspace Descent

Rafael M. Frongillo
Harvard University
raf@cs.berkeley.edu

Mark D. Reid
The Australian National University & NICTA
mark.reid@anu.edu.au

Abstract

We develop a generalization of randomized coordinate descent for smooth convex problems, where the coordinates specify arbitrary subspaces, and derive standard $O(1/\epsilon)$ and $O(1/\log\epsilon)$ rates. For the special case of overlapping 2-block subspaces (i.e. graphs), which has received attention in the literature recently, we derive a convergence rate on a given graph in terms of its algebraic connectivity. Using this connection, we introduce bounds for graph topologies not previously considered. We conclude with preliminary progress toward the interesting open question: what is the best network structure for a given optimization problem?

1 Introduction and Previous Work

Often motivated by large machine learning problems, several randomized coordinate descent methods have appeared in the literature recently, with increasing levels of sophistication. While earlier methods focused on updates which only modified disjoint blocks of coordinates [1, 2], more recent methods allow for more general configurations, such as overlapping blocks [3, 4, 5]. As we will see, there is a common technique underpinning many of the papers mentioned above. Roughly speaking, the recipe is as follows:

1. Derive a quadratic upper bound via Lipschitz continuity;
2. Minimize this upper bound to obtain the update step;
3. Pick a norm based on the update which captures the expected progress per iteration;
4. Use the definition of the dual norm and the convexity of the objective to relate this progress to the optimality gap and a global notion of distance (the function $\mathcal{R}^2$ below);
5. Chain the per-iteration progress bounds into a convergence rate.

We will follow this recipe to present and analyze a general randomized coordinate descent method (Algorithm 1), which we call *randomized subspace descent (RSD)*, whose coordinate updates correspond to m arbitrary subspaces. We will represent each subspace i using an orthogonal projection matrix $\Pi_i \in \mathbb{R}^{n\times n}$, where an update in coordinate i is constrained to be in the image space of Π_i. In other words, if the algorithm performs an update $x^{t+1} \leftarrow x^t + d$, we require $d \in \mathrm{im}(\Pi_i)$.

Formally, our optimization problem is the following:

$$\min_{x\in\mathbb{R}^n} F(x) \quad \text{s.t. } Cx = b\,, \tag{1}$$

for some $C \in \mathbb{R}^{n'\times n}$ and $b \in \mathbb{R}^{n'}$. To achieve the full feasible set $\{x : Cx = b\}$, we choose x^0 with $Cx^0 = b$, and require that the span of the image spaces of $\{\Pi_i\}_{i=1}^m$ is equal to $\mathrm{im}(I - C^+C)$, where C^+ is the Moore-Penrose pseudoinverse.[1] For example, in Section 3 we will need $\sum_i x_i = c$ for some $c \in \mathbb{R}$; by the above, to satisfy this constraint we need only ensure $\mathbb{1} \in \ker\Pi_i$ for all i. Also, in contrast to [1, 2], we do not assume that the matrices Π_i have disjoint images.

[1] By properties of the pseudoinverse, the solutions to $Cx = b$ can be written $x = C^+b + (I - C^+C)y$ for any $y \in \mathbb{R}^n$; taking differences between x^t and x^0 yields the image condition.

We believe that our analysis can be used to recover the smooth-objective results from the papers mentioned above. The previous work dealing with coordinate blocks are immediate special cases, with Π_i being diagonal with value ones on coordinates in the block and zero elsewhere. As we will see in Section 3, graphical settings can be captured by $\Pi_{(i,j)} = \frac{1}{2}(e_i - e_j)(e_i - e_j)^\top$ and more generally, hypergraphs by $\Pi_S = I_S - \frac{1}{|S|}\mathbb{1}_S\mathbb{1}_S^\top$ where $\mathbb{1}_S$ is the binary vector with ones on $S \subseteq [n]$. We note that many of the papers cited above have further generalizations which we ignore for simplicity. For example, the non-Euclidean norms used in Richtárik and Takáč [2] may still be used by invoking their Lemma 10 appropriately; similarly, high-probability results follow from [2, Thm 1]. Results for composite objectives are less immediate, though we believe our approach can be extended to those settings as well.

Finally, we remark that while previous work in coordinate methods was largely motivated by the low per-iteration costs of computing coordinate gradients, it may seem unlikely that such benefits would extend to arbitrary subspaces. As such, we view our contribution as a theoretical unification, providing a simple principled approach to deriving such algorithms in situations where gradient computations in certain subspaces are much more efficient than in the whole space.

2 The Algorithm

We now present our algorithm, randomized subspace descent. The inputs are a smooth convex function $F : \mathbb{R}^n \to \mathbb{R}$, feasible initial point $x^0 \in \mathbb{R}^n$, matrices $\{\Pi_i \in \mathbb{R}^{n\times n}\}_{i=1}^m$ satisfying $\operatorname{span}(\{\Pi_i\}) = \operatorname{im}(I - C^+C)$, smoothness parameters $\{L_i\}_{i=1}^m$, and distribution $p \in \Delta_m$.

Algorithm 1 Randomized Subspace Descent

1: **for** iteration t in $\{0, 1, 2, \cdots\}$ **do**
2: Sample i from p
3: $x^{t+1} \leftarrow x^t - \frac{1}{L_i}\Pi_i \nabla F(x^t)$
4: **end for**

We will assume that F is L_i-smooth with respect to the image space of Π_i; this is step 1 of our recipe. Precisely, we require the existence of constants L_i such that for all $y \in \operatorname{im}(\Pi_i)$,

$$F(x + y) \le F(x) + \langle \nabla F(x), y\rangle + \tfrac{L_i}{2}\|y\|_2^2 \,, \tag{2}$$

and refer to this condition as F being L_i-Π_i-*smooth*. Note that as prescribed by step 2 of our general approach, minimizing this bound over all x' for $y = \Pi_i x'$ yields the update on line 3 of the algorithm by properties of orthogonal projections (see the proof of Theorem 1).

For step 3, we now introduce a seminorm $\|\cdot\|_A$ which will measure the progress per iteration:

$$\|x\|_A := \left(\sum_{i=1}^m \frac{p_i}{L_i}\|\Pi_i x\|_2^2\right)^{1/2} . \tag{3}$$

Note that this is a Euclidean seminorm $\|x\|_A = \langle Ax, x\rangle$ with $A = \sum_i \frac{p_i}{L_i}\Pi_i$. Finally, for step 4 of our recipe, we will need the dual norm of $\|\cdot\|_A$, from which we may define the distance function we need. Let $X(A) := \{x^0 + Ay : y \in \mathbb{R}^n\}$ denote the optimization domain, and $F^{\min} := \min_{x\in X(A)} F(x)$ and $F^{\arg} := \arg\min_{x\in X(A)} F(x)$ denote the minimum and minimizers of F, respectively.

$$\|y\|_A^* := \begin{cases} \langle A^+ y, y\rangle^{1/2} & \text{if } y \in \operatorname{im}(A) \\ \infty & \text{otherwise.} \end{cases} \tag{4}$$

$$\mathcal{R}(x_0) := \max_{x\in X(A):F(x)\le F(x^0)} \max_{x^*\in F^{\arg}} \|x - x^*\|_A^* \,. \tag{5}$$

One can indeed check that $\|\cdot\|_A^*$ is the dual norm of $\|\cdot\|_A$, in the sense that $\left(\frac{1}{2}\|\cdot\|_A^2\right)^* = \frac{1}{2}\|\cdot\|_A^{*2}$.

We are now ready to prove an $O(1/t)$ convergence rate for Algorithm 1. Our analysis borrows heavily from [2] and [4]; we give the proof in Appendix A.

Theorem 1. *Let F, $\{\Pi_i\}_i$, $\{L_i\}_i$, x^0, and p be given as in Algorithm 1, with the condition that F is L_i-Π_i-smooth for all i. Then*

$$\mathbb{E}\left[F(x^t) - F^{\min}\right] \leq \frac{2\mathcal{R}^2(x^0)}{t} \,. \tag{6}$$

As is standard in the literature, when F is strongly convex, we obtain linear convergence.

Theorem 2. *Let F, $\{\Pi_i\}_i$, $\{L_i\}_i$, x^0, and p be given as in Algorithm 1, with the condition that F is L_i-Π_i-smooth for all i, and additionally that F is μ-strongly convex with respect to $\|\cdot\|_A^*$. Then*

$$\mathbb{E}\left[F(x^t) - F^{\min}\right] \leq (1-\mu)^t \left(F(x^0) - F^{\min}\right) . \tag{7}$$

To illustrate the power of Theorem 1, we will show next how to study updates on a graph or hypergraph, each of which correspond to special cases of the subspaces $\{\Pi_i\}_i$.

3 Special Case: Graphs and Hypergraphs

In this section we consider a special case of Algorithm 1, where we have a linear constraint $\sum_i x_i = c$ on the coordinates, and the subspaces correspond to graphs (overlapping pairs of coordinates), or hypergraphs (overlapping subsets of coordinates).[2] In the graphical case, we will leverage existing results in spectral graph theory to analyze new graphs currently not considered in the literature. Note that we focus here on uniform probabilities to highlight the connections to spectral graph theory; for an analysis of the optimal probabilities, see Necoara et al. [4].

Let us first consider an optimization problem on the complete graph, which picks an edge (i,j) uniformly at random and optimizes in coordinates i and j under the constraint that $x_i^{t+1} + x_j^{t+1} = x_i^t + x_j^t$. One can check that this corresponds to the projection matrix $\Pi_{(i,j)} = \frac{1}{2}(e_i - e_j)(e_i - e_j)^\top$, where e_i is the ith standard unit vector. Assuming a global smoothness constant L, one can calculate

$$A = \tfrac{2}{Ln(n-1)} \sum\nolimits_{(i,j)} \Pi_{(i,j)} = \tfrac{1}{L(n-1)} \left(I - \tfrac{1}{n}\mathbb{1}\mathbb{1}^\top\right) \,, \quad A^+ = L(n-1)(I - \tfrac{1}{n}\mathbb{1}\mathbb{1}^\top) \,,$$

where $\mathbb{1}$ is the all-ones vector. Now as $\mathrm{im}(A) = \ker(\mathbb{1})$, this gives

$$\|x\|_A^{*\,2} = L(n-1)\|x\|_2^2 \,. \tag{8}$$

Similarly, the complete rank-k hypergraph gives $\|x\|_A^{*\,2} = L\frac{n-1}{k-1}\|x\|_2^2$. (Compare to eq. (3.10) and the top of p.21 of [4].) Letting $\mathcal{C}_0 = 4L \max_{x \in X(A): F(x) \leq F(x^0)} \max_{x^* \in F^{\arg}} \|x - x^*\|_2^2$, which is independent of the (hyper)graph as long as it is connected, we thus have a convergence rate of $\frac{n-1}{2}\,\mathcal{C}_0\,\frac{1}{t}$ for the complete graph, and more generally $\frac{n-1}{2(k-1)}\,\mathcal{C}_0\,\frac{1}{t}$ for the complete k-graph. Henceforth, we will consider the coefficient in front of $\mathcal{C}_0$ to be the convergence rate.

The above matrix A is a scaled version of what is known as the *graph Laplacian*; given graph G with adjacency matrix $A(G)$ and degree matrix $D(G)$ with the degrees of each vertex on the diagonal, the Laplacian is the matrix

$$\mathcal{L} = \mathcal{L}(G) := D(G) - A(G) \,. \tag{9}$$

One can check that indeed, $\mathcal{L} = 2\sum_{(i,j)\in E(G)} \Pi_{(i,j)}$, meaning $A = \frac{p}{2L}\mathcal{L}$, where $p = 1/|E(G)|$ is the uniform probability on edges.

The graph Laplacian is a well-studied object in spectral graph theory and other domains, and we can use existing results to establish bounds for other graphs of interest. To draw this connection, we note two facts: (1) for symmetric matrices B, the norm $\langle Bx, x\rangle^{1/2}$ can be bounded by the maximum eigenvalue of B, and (2) the maximum eigenvalue of B^+ is equal to the inverse of the smallest nonzero eigenvalue of B, provided again that B is symmetric.[3] Putting these together, we can therefore bound $\|\cdot\|_A^*$ using the smallest nonzero eigenvalue of A, and hence of $\mathcal{L}$. It is easy to see that the smallest eigenvalue is $\lambda_1(G) = 0$ with eigenvector $\mathbb{1}$, and as G is connected, we will

[2] Everything in this section also holds for a graphical or hypergraphical structure on blocks of coordinates; just add Kronecker products with the appropriate identity matrix.

[3] These facts follow from the operator norm and singular-value decomposition for the pseudoinverse, respectively, together with the fact that singular values are eigenvalues for symmetric matrices.

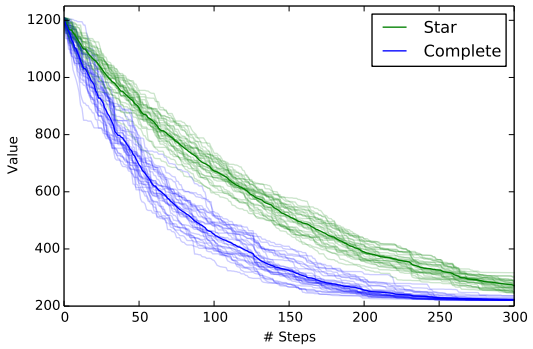

Figure 1: Average (in bold) of 30 runs of a separable objective for the complete and star graphs. The empirical gap in iteration complexity is just under 2 (cf. Fig. 2).

Graph	$\lvert V(G)\rvert$	$\lvert E(G)\rvert$	$\lambda_2(G)$
K_n	n	$n(n-1)/2$	n
P_n	n	$n-1$	$2(1-\cos\frac{\pi}{n})$
C_n	n	n	$2(1-\cos\frac{2\pi}{n})$
$K_{\ell,k}$	$\ell+k$	ℓk	k
B_k	2^k	$k2^{k-1}$	2

Table 1: Algebraic connectivities for common graphs.

have $\lambda_2(G) > 0$. Thus, the smallest nonzero eigenvalue of A is simply $\frac{p}{2L}\lambda_2(G)$, so we have the following for *any* connected graph G:

$$\|x\|_A^{*\,2} \leq 2L\frac{|E(G)|}{\lambda_2(G)}\|x\|_2^2\,. \tag{10}$$

Of course, by the above definition of $\mathcal{C}_0$ and Theorem 1, this yields the result

$$\mathbb{E}\left[F(x^t) - F^{\min}\right] \leq \frac{|E(G)|}{\lambda_2(G)}\mathcal{C}_0\frac{1}{t}\,, \tag{11}$$

showing us how tightly related this eigenvalue is to rate of convergence of Algorithm 1.

The second-smallest eigenvalue $\lambda_2(G)$ is called the *algebraic connectivity* of G, and is itself thoroughly studied in spectral and algebraic graph theory. For example, it is known (and easy to check) that $\lambda_2(K_n) = n$, where K_n denotes the complete graph; this together with $|E(K_n)| = n(n-1)/2$ immediately gives eq. (8). In [6], algebraic connectivities are also given for the path on n vertices P_n, the cycle C_n, the bipartite complete graph $K_{\ell,k}$ for $k < \ell$, and the k-dimensional cube B_k. We collect these eigenvalues together yields Table 1.

Substituting the values in Table 1 into eq. (10), we can directly compare the theoretical convergence rates for different graphs. For example, the star graph $K_{n-1,1}$ has rate $(n-1)(1)/(1) = (n-1)$, which is only a factor of 2 away from the complete graph.[4] The path and cycle fare much worse, yielding roughly $n/2(n^{-2}/2) = n^3$ as n becomes large (applying the Taylor expansion and ignoring π terms). Finally, an interesting result due to Mohar [7] says that for any connected graph on n vertices, we have $\lambda_2(G) \geq 4/(n\mathrm{diam}(G))$ where $\mathrm{diam}(G)$ is the diameter of G. Hence for any connected graph,

$$\mathbb{E}\left[F(x^t) - F^{\min}\right] \leq \frac{n\,|E(G)|\,\mathrm{diam}(G)}{4}\,\mathcal{C}_0\frac{1}{t}\,, \tag{12}$$

which is useful for sparse graphs of small diameter. See Appendix B for more on hypergraphs.

4 Future Work: What is the Optimal Network?

As we have demonstrated above, our general approach to choosing coordinate subspaces combines very naturally with the literature on spectral graph theory, yielding a reasonably rich understanding of the convergence rates for various choices of network structure. In particular, one can use this approach to analyze algorithms for specific networks without needing to start from scratch.

Our study opens up the interesting question: what graph G offers the lowest expected number of iterations? We conjecture that the bound $|E(G)|/\lambda_2(G)$ is minimized by the complete graph.[5] However, in practice if the edges in the graph correspond to physical or logical connections which each incur some cost, it may be desirable to trade off the number of edges with the convergence rate; in this case we expect *expander* graphs, which already have numerous ties to network design [9], to be optimal, as they offer high algebraic connectivity with few edges. Finally, it would be of interest to compute similar bounds for general classes of hypergraphs, to better understand the trade-offs between the convergence rate and the size/connectivity of coordinate subspaces.

[4] While of course these are merely upper bounds on the true rates, they match Figures 1 and 2 quite well.

[5] In particular, a proof seems to follow from results in [8].

References

[1] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[2] 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.

[3] Ion Necoara. Random coordinate descent algorithms for multi-agent convex optimization over networks. *Automatic Control, IEEE Transactions on*, 58(8):2001–2012, 2013.

[4] I Necoara, Y Nesterov, and F Glineur. A random coordinate descent method on large-scale optimization problems with linear constraints. *Technical Report*, 2014.

[5] Sashank Reddi, Ahmed Hefny, Carlton Downey, Avinava Dubey, and Suvrit Sra. Large-scale randomized-coordinate descent methods with non-separable linear constraints. *arXiv preprint arXiv:1409.2617*, 2014.

[6] Nair Maria Maia de Abreu. Old and new results on algebraic connectivity of graphs. *Linear algebra and its applications*, 423(1):53–73, 2007.

[7] Bojan Mohar. The laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications*. Citeseer, 1991.

[8] Sasmita Barik and Sukanta Pati. On algebraic connectivity and spectral integral variations of graphs. *Linear Algebra and its Applications*, 397:209–222, March 2005.

[9] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

A Proofs

Before giving the proof, we note that the result in [2, Thm 11] also holds for general Euclidean norms $\|\cdot\|_{(i)}$. We leave out such extensions as ultimately the only change is in the update step (by leveraging e.g. [2, Lemma 10] instead of our pseudoinverse update) and the form of the dual norm. Lemma 1 below verifies that $\|\cdot\|_A$ is still a seminorm in these cases.

Proof of Theorem 1. To begin, suppose subspace i is chosen at step t, and consider the update $x^{t+1} = x^t - y$ for $y \in \mathrm{im}(\Pi_i)$. The drop in the objective can be bounded using eq. (2),

$$F(x^t) - F(x^t - y) \geq \left\langle \nabla F(x^t), y \right\rangle - \frac{L_i}{2}\|y\|_2^2 \,. \tag{13}$$

By properties of orthogonal projections, we have

$$\operatorname*{arg\,max}_{y \in \mathrm{im}(\Pi_i)} \left\langle \nabla F(x^t), y \right\rangle - \frac{L_i}{2}\|y\|_2^2 \;=\; \operatorname*{arg\,min}_{y \in \mathrm{im}(\Pi_i)} \left\| y - \tfrac{1}{L_i}\nabla F(x^t) \right\|_2 \;=\; \tfrac{1}{L_i}\Pi_i \nabla F(x^t) \,,$$

and choice of y gives our update on line 3. Substituting this y into eq. (13) gives

$$\begin{aligned} F(x^t) - F(x^{t+1}) &\geq \left\langle \nabla F(x^t), \tfrac{1}{L_i}\Pi_i \nabla F(x^t) \right\rangle - \frac{L_i}{2}\|\tfrac{1}{L_i}\Pi_i \nabla F(x^t)\|_2^2 \\ &= \frac{1}{2L_i}\|\Pi_i \nabla F(x^t)\|_2^2 \,. \end{aligned}$$

Now looking at the expected drop in the objective, we have

$$F(x^t) - \mathbb{E}\left[F(x^{t+1})|x^t\right] \geq \sum_{i=1}^m p_i \frac{1}{2L_i}\|\Pi_i \nabla F(x^t)\|_2^2 = \frac{1}{2}\|\nabla F(x^t)\|_A^2 \,. \tag{14}$$

To complete step 4 of our recipe and relate our per-round progress to the gap remaining, we observe that

$$\begin{aligned} F(x^t) - F^{\min} &\leq \max_{x^* \in \arg\min_x F(x)} \left\langle \nabla F(x^t), x^* - x^t \right\rangle \\ &\leq \max_{x^* \in \arg\min_x F(x)} \|\nabla F(x^t)\|_A \, \|x^* - x^t\|_A^* \\ &\leq \|\nabla F(x^t)\|_A \max_{x^* \in \arg\min F} \; \max_{x : F(x) \leq F(x^0)} \|x^* - x\|_A^* \\ &= \|\nabla F(x^t)\|_A \, \mathcal{R}(x^0) \,, \end{aligned}$$

where we used convexity of F, the definition of the dual norm, the fact that $F(x^t)$ is non-increasing in t, and finally the definition of $\mathcal{R}$. We now have $F(x^t) - \mathbb{E}\left[F(x^{t+1})|x^t\right] \geq (F(x^t) - F^{\min})/(2\mathcal{R}^2(x^0))$. The remainder of the proof follows an argument of [4] by analyzing $\Delta_t = \mathbb{E}\left[F(x^t) - F^{\min}\right]$. From the last inequality we have $\Delta_{t+1} \leq \Delta_t - \Delta_t^2/2\mathcal{R}^2(x^0)$, and since $\Delta_{t+1} \leq \Delta_t$, this gives $\Delta_t^{-1} \leq \Delta_{t+1}^{-1} - (2\mathcal{R}^2(x^0))^{-1}$. Summing these inequalities gives the result. □

Proof of Theorem 2. Our proof is essentially that of Nesterov [1, Thm 2] and Richtárik and Takáč [2, Thm 12]. By definition of μ-strongly convex, we have for all $y \in \mathbb{R}^n$,

$$F(y) - F(x^t) \geq \left\langle \nabla F(x^t), y - x^t \right\rangle + \tfrac{\mu}{2}\|y - x^t\|_A^{*\,2} \,.$$

Independently minimizing each side of this inequality over y, we obtain from [2, Lemma 10],

$$F^{\min} - F(x^t) \geq -\tfrac{1}{2\mu}\|\nabla F(x^t)\|_A^2.$$

Now combining with eq. (14), we have

$$F(x^t) - \mathbb{E}\left[F(x^{t+1})|x^t\right] \geq \frac{1}{2}\|\nabla F(x^t)\|_A^2 \geq \mu(F(x^t) - F^{\min}) \,.$$

Taking expectations and rearranging, we have $\mathbb{E}\left[F(x^{t+1}) - F^{\min}\right] \leq (1-\mu)\mathbb{E}\left[F(x^t) - F^{\min}\right]$, from which the result follows by induction. □

Lemma 1. *Let seminorms $\{\|\cdot\|_{(i)}\}_{i=1}^m$ and positive weights $\{w_i\}_{i=1}^m$ be given, and define the function $\|\cdot\|_W : \mathbb{R}^n \to \mathbb{R}$ by*

$$\|x\|_W = \left(\sum_{i=1}^m w_i \|x\|_{(i)}^2\right)^{1/2}. \tag{15}$$

Then $\|\cdot\|_W$ is a seminorm. It is additionally a norm if and only if $\|x\|_{(i)} = 0$ holds for all i only when $x = 0$.

Proof. First, note that we may fold the weights into the seminorms, $\|x\|'_{(i)} := \|\sqrt{w_i}\, x\|_{(i)}$, so we can assume $w_i = 1$ for all i without loss of generality. Let $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ be given by $\varphi(x)_i = \|x\|_{(i)}$. Then $\|x\|_W = \|\varphi(x)\|_2$.

- Absolute homogeneity. First observe that $\varphi(\alpha x) = |\alpha|\varphi(x)$ by homogeneity of the $\|\cdot\|_{(i)}$. Then $\|\alpha x\|_W = \||\alpha|\varphi(x)\|_2 = |\alpha|\|x\|_W$.
- Subadditivity. We first recall the fact that if $x_i \geq y_i$ for all i, then $\|x\|_2 \geq \|y\|_2$. Combining this fact with subadditivity of the $\|\cdot\|_{(i)}$ and then of $\|\cdot\|_2$, we have

$$\begin{aligned}\|x+y\|_W &= \|\varphi(x+y)\|_2 \leq \|\varphi(x) + \varphi(y)\|_2 \\ &\leq \|\varphi(x)\|_2 + \|\varphi(y)\|_2 = \|x\|_W + \|y\|_W.\end{aligned}$$

We now show the norm condition. First, we assume $\|x\|_{(i)} = 0$ for all i implies $x = 0$; we will show Separation. We clearly have $\|0\|_W = 0$. By the above, $\|x\|_W = 0$ implies $\|\varphi(x)\|_2 = 0$, yielding $\|x\|_{(i)} = 0$ for all i by definiteness of $\|\cdot\|_2$, and hence $x = 0$ by assumption.

For the converse, observe that any $x \neq 0$ with $\|x\|_{(i)} = 0$ for all i would imply a violation of definiteness, as $\varphi(x) = 0$ and hence $\|x\|_W = \|\varphi(x)\|_2 = \|0\|_2 = 0$. □

B Hypergraphs

Here we briefly show how to analyze general hypergraphs. Representing a hypergraph as a collection $\mathcal{S}$ of hyperedges $S \subseteq [n]$, we may define the degree matrix $D(\mathcal{S})$ to be the diagonal matrix with $D(\mathcal{S})_{ii} = \#\{S \in \mathcal{S} : i \in S\}$, and the "adjacency" matrix to be $A(\mathcal{S})_{ij} = \sum_{S\in\mathcal{S}: i,j\in S} 1/|S|$. Then for uniform probabilities we have $A = \frac{p}{L}(D(\mathcal{S}) - A(\mathcal{S}))$. This follows from observing that for subset S, we have $\Pi_S = I_S - \frac{1}{|S|}\mathbb{1}_S\mathbb{1}_S^\top$, and counting as we sum. Taking the complete k-graph yields $D(\mathcal{S}) = \binom{n-1}{k-1} I$ and $A(\mathcal{S})_{ij} = \frac{1}{k}\binom{n-2}{k-2} = \frac{k-1}{k(n-1)}\binom{n-1}{k-1}$ for $i \neq j$ and $A(\mathcal{S})_{ii} = \frac{1}{k}\binom{n-1}{k-1}$; putting these together gives $A = \frac{1}{L\binom{n}{k}}\frac{n}{k}\frac{k-1}{n-1}\binom{n-1}{k-1}(I - \frac{1}{n}\mathbb{1}\mathbb{1}^\top) = \frac{n-1}{L(k-1)}(I - \frac{1}{n}\mathbb{1}\mathbb{1}^\top)$. Similar computations may be done for other hypergraphs of interest.

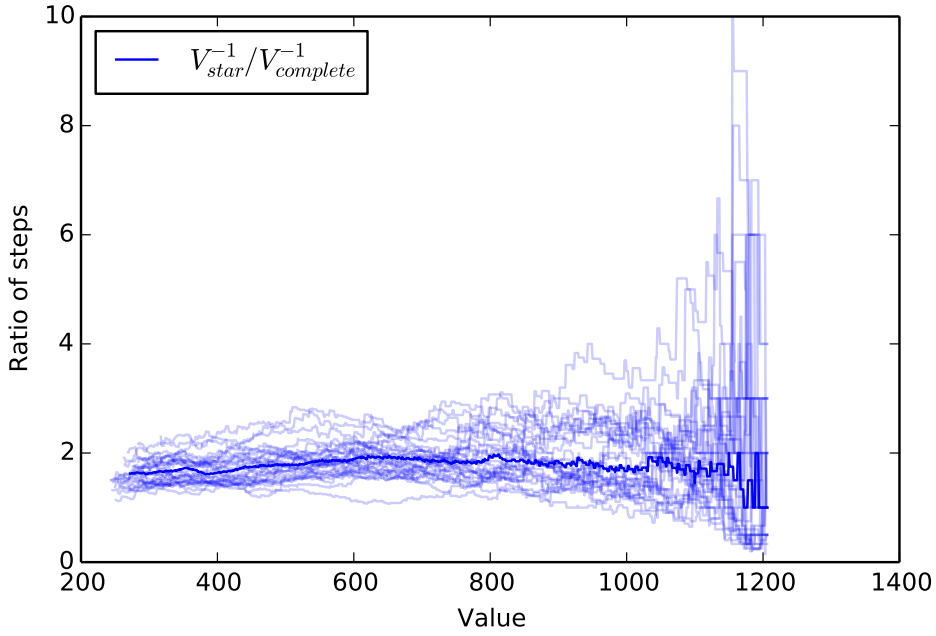

Figure 2: Thirty runs of a separable objective under the complete and star graphs. The ratio between star and complete of the number of iterations needed to achieve a given objective value is plotted, with the average in bold.