Coordinate Descent without Coordinates: Tangent Subspace Descent on Riemannian Manifolds

Nam Ho-Nguyen

Discipline of Business Analytics
The University of Sydney

Variational Analysis and Optimization Webinar
February 10, 2021

Authors

David Huckleberry Gutman
Industrial, Manufacturing and Systems Engineering
Texas Tech University
david.gutman@ttu.edu

Nam Ho-Nguyen
Discipline of Business Analytics
The University of Sydney
nam.ho-nguyen@sydney.edu.au

Supported, in part, by Award N660011824020 from the DARPA Lagrange Program and NSF Award 1740707.

Nonlinear optimization

Problem

$$\min_{x \in C} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $C \subseteq \mathbb{R}^n$. (Often both convex, but non-convex becoming more common.)

Applications:

- *Economics/finance*:
  - Portfolio and risk optimization.
  - Planning/production.
- *Engineering*:
  - Control.
  - Circuit/structural design.
  - Signal/image processing.
- *Statistics and machine learning*:
  - Data fitting: classification, regression, matrix completion.
  - Design of experiments.

Numerical methods for nonlinear optimization

- **The Early Era**: pre–1980s
  - First-order methods: gradient descent, Frank-Wolfe, perceptron.
- **The Medium Scale Era**: 1980s–2000s
  - Interior-point & other second order methods
  - Conic programming (second-order cone, semidefinite)
  - Strong theory & industry ready software packages with great accuracy
  - Elaborate algorithms (involving matrix inversion) for generic problems
- **The Large Scale Era**: 2000s–now
  - Lots of data $\implies$ large-scale problems
  - Goal: modest accuracy & cheap $O(n)$ iterations
  - Resurgence of **first-order methods**
  - Simple algorithms (matrix inversion-free).

Disclaimer: this does not include progress on **discrete optimization** methods.

Gradient descent and extensions

Consider the problem

$$f^* := \min_{x \in \mathbb{R}^n} f(x).$$

- **Gradient descent:** f differentiable

$$x_{k+1} = x_k - t_k \nabla f(x_k).$$

- **Proximal gradient:** $f = g + h$, g differentiable

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ g(x_k) + \langle \nabla g(x_k), x - x_k \rangle + \frac{t_k^2}{2} \|x - x_k\|_2^2 + h(x) \right\}.$$

- **Stochastic gradient descent:** $f(x) = \mathbb{E}[g(x;\xi)]$, $\xi \sim \mathbb{P}$

$$x_{k+1} = x_k - t_k \nabla_x g(x_k, \xi_k), \quad \xi_k \sim \mathbb{P}.$$

Coordinate descent in $\mathbb{R}^n$

Several key problems have n being **very** large. In which case, we run **coordinate descent** (CD) [Beck and Tetruashvili, 2013, Nesterov, 2012]:

- Given x_k, do the following: (cyclic CD)

$$
\begin{aligned}
&\text{Set } y_{k,0} = x_k \\
&\text{for } j \in \{1, \ldots, n\} \\
&\qquad y_{k,j} = y_{k,j-1} - t_{k,j} e_j \underbrace{e_j^\top \nabla f(y_{k,j-1})}_{=\partial_j f(y_{k,j-1})} \\
&\text{Set } x_{k+1} = y_{k,n}.
\end{aligned}
$$

(We can also randomly pick the index $j \implies$ randomized CD).

Optimization on manifolds

Manifold domains

$$\min_{x \in M} f(x), \quad \text{where } M \text{ is a } \textbf{manifold}.$$

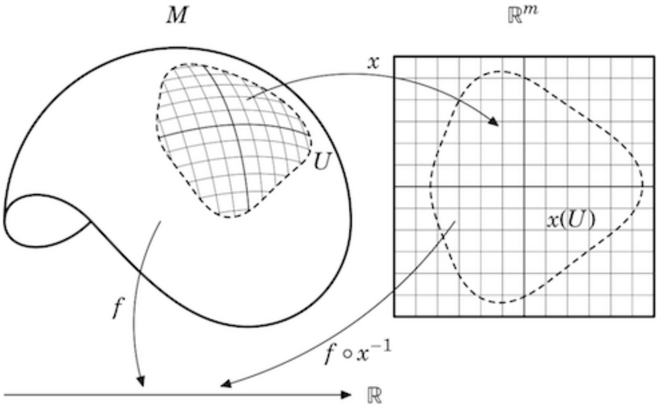
(Informally) M is "locally flat" but (possibly) globally curved set of dimension n on which we can do calculus.

- Think about Earth: looks flat from local perspective, but globally curved.

Smooth n-Manifold (Formal): topological space M that's

- *Locally Euclidean and Smooth*: Every point $x \in M$ has neighborhood U homeomorphic to open set in $\mathbb{R}^n$.
  - This means locally flat
- *Smooth Compatibility*: Local Euclidean homeomorphisms are smoothly compatible.
  - Technical, but allows us to do calculus.

Local flatness of manifolds



Modelling via manifolds

Question

What is the benefit of modelling the domain as a manifold?

Answer:

1. Some domains have **structural symmetry and/or invariance**, and manifolds are primed to capture various geometric aspects of the domain.
2. Some problems are non-convex, but modelling the domain as a manifold M and endowing M with an appropriate **Riemannian metric** makes them **geodesically convex** (defined later).

Modelling via manifolds

Manifolds arise in several important applications.

- **Principal component analysis.** Suppose we have points $u_1, \ldots, u_K \in \mathbb{R}^n$ with zero mean. Find a p-**dimensional subspace** of $\mathbb{R}^n$ to project the points onto which preserves the variance (as much as possible):

$$\max_{X \in M} \frac{1}{K} \sum_{k \in [K]} \|XX^\top u_k\|_2^2, \quad M := \mathrm{St}(p, n) = \left\{X \in \mathbb{R}^{n \times p} : X^\top X = I_p\right\}.$$

  Advantages: manifold optimization algorithms implicitly handle the $X^\top X = I_p$ constraints by exploiting the structure of the **Stiefel manifold** domain.

Modelling via manifolds

Manifolds arise in several important applications.

- **Low rank matrix approximation for recommender systems.** We have n customers and k products. Matrix $U \in \mathbb{R}^{n\times k}$ captures ratings, but we only see a few. How can predict unobserved ratings?

$$\min_{X\in M} \sum_{(i,j)\in\Omega\subset[n]\times[k]} (U_{ij} - X_{ij})^2, \quad M := \left\{ X \in \mathbb{R}^{n\times k} : \begin{array}{r} X = C^\top P \\ C \in \mathbb{R}^{m\times n} \\ P \in \mathbb{R}^{m\times k} \end{array} \right\}$$

  The hypothesis is that each customer i has an attribute vector $c_i \in \mathbb{R}^m$, each product j has an attribute vector $p_j \in \mathbb{R}^m$, then the rating is

$$u_{ij} = c_i^\top p_j.$$

Modelling via manifolds

Manifolds arise in several important applications:

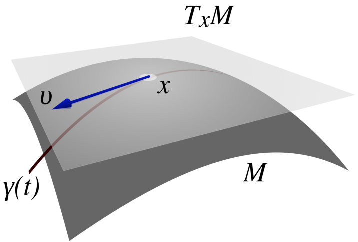
- **Gaussian Fisher-Rao distance.** The family of non-degenerate zero-mean Gaussians $N(0, \Sigma)$ can be parametrized by positive definite matrices $\Sigma \in \mathbb{S}^n_{++}$. The **Fisher-Rao distance** is

$$d(\Sigma_0, \Sigma_1) = \frac{1}{\sqrt{2}} \left\| \log \left(\Sigma_1^{-1/2} \Sigma_0 \Sigma_1^{-1/2} \right) \right\|_F .$$

  This has been used in statistical estimation and information geometry.
  The Fisher-Rao distance is **non-convex** in the Euclidean geometry, but becomes **geodesically convex** when $\mathbb{S}^n_{++}$ is endowed with its **intrinsic metric**.

First-order methods on manifolds

Concept	General manifold M	$M = \mathbb{R}^n$
directions from $x \in M$	tangent space $T_x M \cong \mathbb{R}^n$	$T_x M \equiv \mathbb{R}^n$
gradients of f	$\nabla f(x) \in T_x M$	$\nabla f(x) \in \mathbb{R}^n$
Riemannian metric	$\langle \cdot, \cdot \rangle_x$ for each $T_x M$	usual inner product
comparing $T_x M$ vs $T_y M$	$\Gamma_x^y : T_x M \to T_y M$	$\Gamma_x^y = I_n$
Movement in a direction	$v \in T_x M \mapsto \mathrm{Exp}_x(v) \in M$	$\mathrm{Exp}_x(v) = x + v$
distance x, $y = \mathrm{Exp}_x(v)$	$d(x, y) = \|v\|_x = \sqrt{\langle v, v \rangle_x}$	$d(x, y) = \|v\|_2$



A function $f : M \to \mathbb{R}$ is **geodesically convex** if $t \mapsto f(\mathrm{Exp}_x(tv))$ is convex in $\mathbb{R}$ for any $x \in M$, $v \in T_x M$.

Example: positive definite matrices

Consider $M = \mathbb{S}^n_{++}$ [Sra and Hosseini, 2015].

- **Tangent space:** For $X \in M$, $T_X M = \mathbb{S}^n$.
- **Riemannian metric:** Given $X \in T_X M$ and $V_1, V_2 \in T_X M$, define

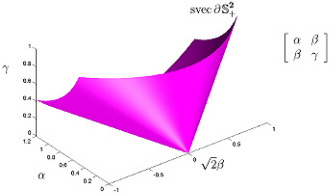
$$\langle V_1, V_2 \rangle_X = \mathrm{Tr}(V_1 X^{-1} V_2 X^{-1}).$$

- **Parallel transport:** Given $X, Y \in M$ and $V \in T_X M$,

$$\Gamma_X^Y(V) = (YX^{-1})^{1/2} V (X^{-1} Y)^{1/2} \in T_Y M.$$

- **Exponential map:** Given $X \in M$, $V \in T_X M$,

$$\mathrm{Exp}_X(V) = X^{1/2}\, \mathrm{Expm}(X^{-1/2} V X^{-1/2}) X^{1/2}.$$



Riemannian Gradient Descent

Consider *unconstrained* optimization on smooth n-manifold

$$\min_{x \in M} f(x)$$

with $f : M \to \mathbb{R}$ differentiable.

Riemannian Gradient Descent:

$$x_{k+1} = \mathrm{Exp}_{x_k}(-t_k \nabla f(x_k))$$
$$(x_{k+1} = x_k - t_k \nabla f(x_k) \quad \text{Euclidean})$$

Convergence Rates [Zhang and Sra, 2016]: Under suitable conditions on M and adaption of L-Lipschitz ∇f, if $t_k = \frac{1}{L}$ then

- $\min_{i=0,\dots,k} \|\nabla f(x_i)\|_{x_i} = O\left(\frac{1}{\sqrt{k}}\right)$ ...
- but if f geodesically convex $\Rightarrow f(x_k) - f^* = O\left(\frac{1}{k}\right)$

Subspace descent

Consider the extension of CD to **subspace descent**:

- Pick subspaces $\{S_j\}_{j\in[m]}$ such that $\text{Span}\left(\bigcup_{j\in[m]} S_j\right) = \mathbb{R}^n$.
- Given x_k, do the following:

$$
\begin{aligned}
&\text{Set } y_{k,0} = x_k \\
&\text{for } j \in \{1, \ldots, m\} \\
&\qquad y_{k,j} = y_{k,j-1} - t_j P_{S_j} \nabla f(y_{k,j-1}) \\
&\text{Set } x_{k+1} = y_{k,m}
\end{aligned}
$$

(Randomized version established by Frongillo and Reid [2015].)

Tangent subspace descent

Generalization: subspaces of $\mathbb{R}^n \implies$ subspaces of $T_x M$. We call this **tangent subspace descent (TSD)**.

Set $y_{k,0} = x_k$
for $j \in \{1, \ldots, m\}$
  Pick a subspace $S_{k,j} \subset T_{y_{k,j-1}} M$
  $y_{k,j} = \mathrm{Exp}_{y_{k,j-1}} \left(-t_j P_{S_{k,j}} \nabla f(y_{k,j-1})\right)$
Set $x_{k+1} = y_{k,m}$

- In $\mathbb{R}^n$, the subspaces remain the same.
- On a general M, $S_{k,1}, \ldots, S_{k,m}$ belong to different vector spaces $T_{y_{k,0}} M, \ldots, T_{y_{k,m-1}} M$.
- How should we pick the subspaces?

The problem of subspace selection

If we choose $S_{k,1}, \ldots, S_{k,m}$ poorly, then we may not converge.

Theorem

There exists a subspace selection rule and constant $\epsilon > 0$ such that $f(x_k) > \epsilon$ for all k.

Proof idea.

Take $M = \mathbb{R}^n$, $f(x) = \frac{1}{2}\|x\|_2^2$, $\mathrm{Exp}_x(v) = x + v$. Choose $\epsilon = \|x_0\|_2/4$. For $k \geq 1$, choose

$$S_{k,j} = \mathrm{Span}\left(\{v_{k,j}\}\right), \quad \|v_{k,j}\|_2 = 1,\ \langle v_{k,j}, y_{k,j-1}\rangle = \sqrt{(f(x_{k-1}) - \epsilon)/m},$$

and $y_{k,j} = y_{k,j-1} - \langle v_{k,j}, y_{k,j-1}\rangle v_{k,j}$. □

Subspace selection criterion

Lemma (Sufficient decrease leads to convergence)

Suppose there exists $\eta, \eta' > 0$ *such that the outer iterates* $\{x_k\}$ *satisfy*

$$f(x_k) - f(x_{k+1}) \geq \eta \|\nabla f(x_k)\|_{x_k}^2 \quad \text{or} \quad f(x_k) - f(x_{k+1}) \geq \eta'.$$

Then under suitable regularity conditions on M *and* f

- $\min_{i=0,\ldots,k} \|\nabla f(x_i)\|_{x_i} = O\left(\frac{m}{\sqrt{k}}\right)$.
- *if* f *geodesically convex* $\Rightarrow f(x_k) - f^* = O\left(\frac{m^2}{k}\right)$.

Lemma (Sufficient decrease)

For appropriately chosen step sizes, there exists $C > 0$ *such that we have*

$$f(x_k) - f(x_{k+1}) \geq C \sum_{j \in [m]} \left\| P_{S_{k,j}} \nabla f(y_{k,j-1}) \right\|_{y_{k,j-1}}^2.$$

Key assumption

Assumption (When inner iterates are close, the subspaces are close to orthogonal)

There exists $r > 0$, $\gamma \in [0, 1]$ such that for any outer iterate $k \geq 1$ the subspaces $\{S_{k,j}\}_{j\in[m]}$ are chosen to generate the inner iterates $y_{k,j}$, $j \in [m]$ so that

there exists an orthogonal decomposition $\{D_{k,j}\}_{j\in[m]}$ of $T_{x_k}M$ *such that*

$$\max_{j\in[m]} d(x_k, y_{k,j}) < r \implies \left\| \Gamma_{y_{k,j-1}}^{x_k} P_{S_{k,j}} - P_{D_{k,j}} \right\|_{x_k} \leq \gamma.$$

(We use the induced operator norm from $\langle \cdot, \cdot \rangle_{x_k}$.)

When does the assumption hold?

- When $M = \mathbb{R}^n$ and the subspace decomposition $\{S_j\}_{j\in[m]}$ is fixed throughout.
- When M is a product manifolds $M = M^1 \times \cdots \times M^m$. Then for $x = (x^1, \ldots, x^m) \in M$, $T_x M \cong T_{x^1} M^1 \oplus \cdots \oplus T_{x^m} M^m$. Take the subspaces as
$$S_{k,j} = T_{y^j_{k,j-1}} M.$$
- For general M, fix an orthogonal decomposition $\{D^k_j\}_{j\in[m]}$ of $T_{x_k} M$, and at step j of iteration k, we parallel transport it to $T_{y_{k,j-1}} M$:
$$S_{k,j} = \Gamma_{x_k}^{y_{k,j-1}} D^k_j.$$

Convergence result

Lemma

Suppose the assumption holds. Then there exists $\eta, \eta' > 0$ such that

$$\sum_{j\in[m]} d(y_{k,j-1}, y_{k,j}) \le r \implies f(x^k) - f(x^{k+1}) \ge \eta \|\nabla f(x^k)\|_{x^k}^2$$

$$\sum_{j\in[m]} d(y_{k,j-1}, y_{k,j}) > r \implies f(x^k) - f(x^{k+1}) \ge \eta'.$$

Theorem

Suppose the assumption holds, then under suitable regularity conditions on M and f

- $\min_{i=0,\ldots,k} \|\nabla f(x_i)\|_{x_i} = O\left(\frac{m}{\sqrt{k}}\right)$.
- *if f geodesically convex* $\Rightarrow f(x_k) - f^* = O\left(\frac{m^2}{k}\right)$.

Orthogonal matrices

A non-trivial example where the assumption holds: **orthogonal matrices** (see Edelman et al. [1998])

$$
\begin{aligned}
M := O_n &= \left\{ Y \in \mathbb{R}^{n\times n} : Y^\top Y = YY^\top = I_n \right\} \\
T_Y M &= \left\{ YA : A \in \mathbb{R}^{n\times n},\ A = -A^\top \in \mathrm{Skew}_n \right\} \\
\langle YA, YB \rangle_Y &= \mathrm{Tr}(A^\top B) \\
\mathrm{Exp}_Y(YA) &= Y \,\mathrm{Expm}(A).
\end{aligned}
$$

Given $YA \in T_Y M$ and $Z = \mathrm{Exp}_Y(YC)$, parallel transport of YA from $T_Y M$ to $T_Z M$ is

$$\Gamma_Y^Z(YA) = Z\,\mathrm{Expm}(C/2)^\top A\,\mathrm{Expm}(C/2).$$

Orthogonal matrices

Fix $X_k = Y_{k,0}$. Then $T_{X_k} M = \{X_k A : A \in \mathrm{Skew}_n\}$.

An orthonormal basis for $T_{X_k} M$ is

$$\left\{ \frac{1}{\sqrt{2}} X_k (e_i e_l^\top - e_l e_i^\top) : 1 \le i < l \le n \right\}.$$

Let $m = n(n-1)/2$, order the (i, l) indices as $(i_1, l_1), \ldots, (i_m, l_m)$.

Subspace selection rule for O_n

Pick

$$S_{k,j} = \mathrm{Span}\left(Y_{k,j-1}(e_{i_j} e_{l_j}^\top - e_{l_j} e_{i_j}^\top) \right) \subset T_{Y_{k,j-1}} M.$$

Orthogonal matrices

Let $C_{k,j} \in \text{Skew}_n$ be such that $\text{Exp}_{Y_{k,j-1}}(Y_{k,j-1} C_{k,j}) = X_k$. Parallel transporting the subspaces

$$S_{k,j} = \text{Span}\left(Y_{k,j-1}(e_{i_j} e_{l_j}^\top - e_{l_j} e_{i_j}^\top)\right) \subset T_{Y_{k,j-1}} M$$

back to $T_{X_k} M$ we have a set

$$\left\{X_k \,\text{Expm}(C_{k,j}/2)^\top (e_{i_j} e_{l_j}^\top - e_{l_j} e_{i_j}^\top) \,\text{Expm}(C_{k,j}/2) : j \in [m]\right\}.$$

To prove the assumption: show that when $C_{k,j}$ are small, then the set is "close" to the orthogonal decomposition

$$\left\{X_k (e_{i_j} e_{l_j}^\top - e_{l_j} e_{i_j}^\top) : j \in [m]\right\} \subset T_{X_k} M.$$

Randomized TSD

Pick a subspace decomposition $\{S_k(\xi)\}_{\xi\in\Xi}$ of $T_{x_k}M$

Sample $S_k(\xi)$ at random, $\xi \sim \mathbb{P}$

Set $x_{k+1} = \mathrm{Exp}_{x_k}\left(-t_k P_{S_k(\xi)} \nabla f(x_k)\right)$.

Lemma (Randomized sufficient decrease)

For appropriately chosen step sizes, there exists $C > 0$ such that we have

$$f(x_k) - \mathbb{E}[f(x_{k+1}) \mid x_k] \geq C \cdot \mathbb{E}\left[\left\|P_{S_k(\xi)} \nabla f(x_k)\right\|_{x_k}^2 \mid x_k\right].$$

Convergence

Assumption

There exists $\eta > 0$ such that, for all x, we can construct a subspace decomposition $\{S_x(\xi)\}_{\xi \in \Xi}$ and distribution $\xi \sim \mathbb{P}$ which satisfies

$$\mathbb{E}_{\xi \sim \mathbb{P}} \left[\left\| P_{S_x(\xi)} v \right\|_x^2 \right] \geq \eta \|v\|_x^2$$

for any $v \in T_x M$.

Theorem

If the assumption holds, then under suitable regularity conditions on M and f

- $\min_{i=0,\ldots,k} \mathbb{E}\left[\|\nabla f(x_i)\|_{x_i}\right] = O\left(\frac{m}{\sqrt{k}}\right)$.
- *if f geodesically convex* $\Rightarrow \mathbb{E}[f(x_k)] - f^* = O\left(\frac{m^2}{k}\right)$.

A randomized scheme for the Stiefel manifold

$$M := \mathrm{St}(p,n) = \left\{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\right\}$$

$$T_X M = \left\{ XA + \sum_{l\in[p]} b_l e_l^\top : \begin{array}{c} A = -A^\top \in \mathrm{Skew}_p \\ X^\top b_l = 0 \quad \forall l \in [p] \end{array} \right\}.$$

Randomized selection rule

For $T_X M$

With probability $1/(p(p-1))$: $X(e_i e_l^\top - e_l e_i^\top)$

With probability $1/(2p)$: $(I_n - XX^\top) z_l e_l^\top, \quad z_l \sim N(0, I_n)$.

Theorem

With the above randomization scheme,

$$\mathbb{E}_{\xi\sim\mathbb{P}}\left[\left\|P_{S_x(\xi)} v\right\|_x^2\right] \geq \eta \|v\|_x^2, \quad \eta = \min\left\{\frac{1}{p(p-1)}, \frac{1}{2p(n-p)}\right\}.$$

Preliminary numerical study

We test deterministic TSD on linear optimization problems in O_n:

$$\min_{Y \in O_n} \mathrm{Tr}(D^\top Y).$$

We benchmarked against Riemannian gradient descent.

- We cycle through the basis $\{ Y(e_i e_l^\top - e_l e_i^\top) : 1 \le i < l \le n \}$.
- This allows efficient computation of the matrix exponential and exact step size selection.
- Random instances were generated for $n = 50, 100, 150, 200$.

Results

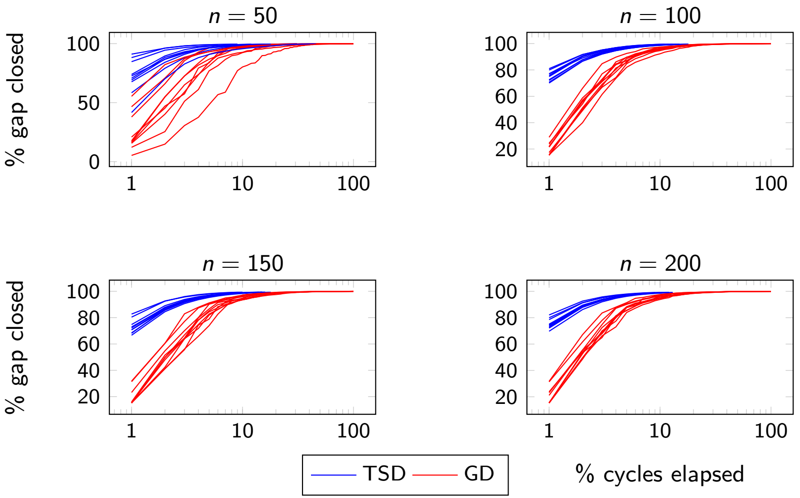


Figure: Results for TSD (blue) vs GD (red). Horizontal axis: cycles elapsed as a percentage of the largest number of cycles for that instance. Vertical axis: gap closed as a percentage of the best objective value found across both algorithms.

Conclusions and future work

Contributions:

- An analogy of coordinate descent to Riemannian manifolds: **tangent subspace descent**.
- Counterexamples and sufficient conditions for subspace selection rules.
- Convergence guarantees for geodesically convex and nonconvex functions.
- Specific subspace selection rules for Stiefel manifolds.

Future work:

- Schemes for different types of manifolds.
- Proximal setting for composite (smooth + nonsmooth) problems.
- Finite-sum problems.

Paper: `https://arxiv.org/abs/1912.10627`.

A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013. doi: 10.1137/120887679.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954.

R. Frongillo and M. D. Reid. Convergence analysis of prediction markets via randomized subspace descent. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3034–3042. Curran Associates, Inc., 2015.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001.

S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015. doi: 10.1137/140978168.

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.