Supplement to
"A Semismooth Newton Method for Fast, Generic Convex Programming"

Alnur Ali [*1] **Eric Wong** [*1] **J. Zico Kolter** [2]

This document contains proofs and supplementary details for the paper "A Semismooth Newton Method for Fast, Generic Convex Programming". All section, equation, and figure numbers in this supplementary document are preceded by the letter S (all numbering without an S refers to the main paper).

S.1. Proof of Lemma 3.2

The proof relies on the proof of Lemma 3.6, below. Let $z, \delta \in \mathbf{R}^3$, and let $\delta \to 0$. Suppose $z + \delta$ converges to a point that falls into one of the first three cases given in Section 2. Then, from the statement and proof of Lemma 3.6, an element $J_{P_{\mathcal{K}^*_{\text{exp}}}}(z+\delta)$ of the generalized Jacobian of the projection onto the dual of the exponential cone at $z+\delta$, is just a matrix with fixed entries, since projections onto convex sets are continuous. If $z+\delta$ converges to a point that falls into the fourth case, then brute force, *e.g.*, using symbolic manipulation software, reveals that an element of the generalized Jacobian (*i.e.*, the inverse of the specific 4x4 matrix D given in (S.6), below) is also a constant matrix, even as $z_1^\star, z_2^\star, \nu^\star \to 0$; for completeness, we give D^{-1} in (S.26), at the end of the supplement. Thus in all the cases, the Jacobian is a constant matrix, which is enough to establish that the limit in (15) exists. □

S.2. Proof of Lemma 3.3

First, we give a useful result; its proof is elementary.

Lemma S.2.1. *The affine transformation, $AF+b$, of a (strongly) semismooth map $F : \mathbf{R}^k \to \mathbf{R}^k$, with $A \in \mathbf{R}^{k\times k},\ b \in \mathbf{R}^k$, is (strongly) semismooth.*

Proof. First of all, we have that a map $F : \mathbf{R}^k \to \mathbf{R}^k$ is (strongly) semismooth if and only if its components F_i, for $i = 1, \ldots, k$, are (strongly) semismooth (Qi & Sun, 1993, Corollary 2.4). Additionally, we have that (strongly) semismooth maps are closed under linear combinations (Izmailov & Solodov, 2014, Proposition 1.75). Putting the two pieces together gives the claim. □

Now, from Lemma 3.1, we have that the projections onto the nonnegative orthant, second-order cone, positive semidefinite cone, as well as the free cone (an affine map, hence strongly semismooth (Facchinei & Pang, 2007, Proposition 7.4.7)), are all strongly semismooth. The map F, defined in (16), is just an affine transformation of these projections; thus, by (S.2.1), it is strongly semismooth.

When $\mathcal{K}$, from (2), is the exponential cone, the analogous claim that the map F is semismooth follows, from Lemma 3.2, in a similar way. □

[*]Equal contribution [1]Machine Learning Department, Carnegie Mellon University [2]Computer Science Department, Carnegie Mellon University. Correspondence to: Alnur Ali <alnurali@cmu.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

S.3. Proof of Lemma 3.4

Proof. We have that (i) the projection onto a convex set (*e.g.*, the nonnegative orthant, second-order cone, positive semidefinite cone, exponential cone, and free cone), naturally, yields a convex set; (ii) the affine image of a convex set is a convex set; and (iii) retaining only some of the coordinates of a convex set is a convex set (Boyd & Vandenberghe, 2004, page 38). Hence, the components F_i, for $i = 1, \ldots, 3k$, of the map $F : \mathbf{R}^{3k} \to \mathbf{R}^{3k}$, defined in (16), are convex functions. Thus, by Clarke (1990, Proposition 1.2), the ith row of any element of the generalized Jacobian is just a subgradient of F_i. Now observe that the element J of the generalized Jacobian, given in (17), is given by finding subgradients of the F_i. □

S.4. Jacobian of the projection onto the second-order cone

In Section 3.2, we stated that, in one case, the Jacobian of the projection onto the second-order cone at some point $z = (z_1, z_2) \in \mathbf{R}^m$, with $z_1 \in \mathbf{R}^{m-1},\ z_2 \in \mathbf{R}$, is a low-rank matrix $D \in \mathbf{R}^{m \times m}$; the matrix D is given by

$$D = \begin{bmatrix} \frac{1}{2} + \frac{z_2}{2\|z_1\|_2} - \frac{z_2}{2}\frac{(z_1)_1^2}{\|z_1\|_2^3} & -\frac{z_2}{2}\frac{(z_1)_1(z_1)_2}{\|z_1\|_2^3} & \cdots & \frac{1}{2}\frac{(z_1)_1}{\|z_1\|_2} \\ -\frac{z_2}{2}\frac{(z_1)_1(z_1)_2}{\|z_1\|_2^3} & \frac{1}{2} + \frac{z_2}{2\|z_1\|_2} - \frac{z_2}{2}\frac{(z_1)_2^2}{\|z_1\|_2^3} & \cdots & \frac{1}{2}\frac{(z_1)_2}{\|z_1\|_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\frac{(z_1)_1}{\|z_1\|_2} & \frac{1}{2}\frac{(z_1)_2}{\|z_1\|_2} & \cdots & \frac{1}{2} \end{bmatrix}, \tag{S.1}$$

which can be seen as the sum of diagonal and low-rank matrices. Here, $(z_1)_i$ denotes the ith component of z_1.

S.5. Proof of Lemma 3.5

Rewrite the projection onto the positive semidefinite cone as (11) as $P_{\mathcal{K}_{\text{psd}}}(Z) = Q \max(\Lambda, 0) Q^T$, where $Z = Q \max(\Lambda, 0) Q^T$ is the eigenvalue decomposition of some real, symmetric matrix Z, and the $\max$ here is interpreted diagonally. Then, using the chain rule (Magnus & Neudecker, 1995), we get that

$$\begin{aligned} J_{P_{\mathcal{K}_{\text{psd}}}}(\mathbf{vec}\, Z)(d\, \mathbf{vec}\, Z) &= d\, \mathbf{vec}\, P_{\mathcal{K}_{\text{psd}}}(Z) \\ &= \mathbf{vec}\left((dQ)\max(\Lambda, 0)Q^T + Q(d\max(\Lambda, 0))Q^T + Q\max(\Lambda, 0)(dQ)^T\right); \end{aligned}$$

so, what remains is computing (each column of) dQ and $d\max(\Lambda, 0)$, *i.e.*, the differential of (each column of) the matrix of eigenvectors, and the differential of $\max(\Lambda, 0)$, respectively. From Magnus & Neudecker (1995, Chapter 8), we get that

$$dQ_i = (\Lambda_{ii} I - Z)^+ (dZ) Q_i,$$

where Z^+ denotes the pseudo-inverse of the matrix Z, and that

$$[d\max(\Lambda, 0)]_{ii} = I_+(\Lambda_{ii}) Q_i^T (dZ) Q_i,$$

by applying the chain rule; here, $I_+(\cdot)$ is the indicator function of the nonnegative orthant, *i.e.*, it equals 1 if its argument is nonnegative and 0 otherwise. Replacing dZ with some real, symmetric matrix $\tilde{Z}$ yields the claim. □

S.6. Further details on the per-iteration costs of SCS, Newton-ADMM, and CVXOPT

Here, we elaborate on the costs of a single iteration of SCS, Newton-ADMM, and CVXOPT. For simplicity, we consider the case where the cone $\mathcal{K}$, in the cone program (1), is just a single cone (handling the case where $\mathcal{K}$ is the direct product of multiple cones is not hard); also, we are mostly interested in the high-dimensional case, where $n > m$.

During a single iteration of SCS, described in (6) – (8), we must carry out the computations outlined below:

- We must update the $\tilde{u}$ variable, which costs $O(\max\{n^2, m^2\})$ (see Section 4.1 of O'Donoghue et al. (2016)).
- We must update the u variable, the cost of which is dominated by the cost of projecting an m-vector onto the dual cone $\mathcal{K}^*$; for the case of projecting onto the positive semidefinite cone, we equivalently consider a matrix with dimensions $\sqrt{m} \times \sqrt{m}$. These costs are as follows:

  - For the nonnegative orthant, $\mathcal{K}_{\text{no}}$, the cost is $O(m)$.
  - For the second-order cone, $\mathcal{K}_{\text{soc}}$, the cost is $O(m)$.
  - For the positive semidefinite cone, $\mathcal{K}_{\text{psd}}$, the cost is $O(m^{3/2})$.
  - For the exponential cone, $\mathcal{K}_{\text{exp}}$, the cost is roughly $O(m^3)$.

- We must update the v variable, which has negligible cost.

Summing up, the cost of a single iteration of SCS is $O(\max\{n^2, m^2\})$ plus the cost of projecting onto the dual cone $\mathcal{K}^*$, as claimed in the main paper.

For Newton-ADMM, we must compute the ingredients on both sides of (19), F and J, as well as run GMRES and the backtracking line search. Computing both F and J can be seen as essentially costing the same as a single iteration of SCS, *i.e.*, the cost of projecting onto the dual cone $\mathcal{K}^*$ plus $O(\max\{n^2, m^2\})$; the backtracking line search, then, costs the number of backtracking iterations times the aforementioned cost. Furthermore, running GMRES costs $O(\max\{n^2, m^2\})$, assuming it returns early. Hence the cost of a single iteration of Newton-ADMM is (as claimed in the main paper) the number of backtracking iterations times the sum of two costs: the cost of projecting onto the dual cone $\mathcal{K}^*$ plus $O(\max\{n^2, m^2\})$.

Finally, turning to the interior-point method CVXOPT, it can be seen that the per-iteration cost here is dominated by solving the Newton system (1.11) in Andersen et al. (2011), essentially costing $O(n^3)$.

We mention that the above per-iteration costs can, of course, be improved by taking advantage of sparsity.

S.7. Proof of Lemma 3.6

First, from the Moreau decomposition given in (13), we get that

$$J_{P_{\mathcal{K}^*_{\text{exp}}}}(z) = I - J_{P_{\mathcal{K}_{\text{exp}}}}(-z);$$

so, what remains is to compute $J_{P_{\mathcal{K}_{\text{exp}}}}(z)$, for some $z \in \mathbf{R}^m$. Looking back at the first three cases given in Section 2, we get that

$$J_{P_{\mathcal{K}_{\text{exp}}}}(z) = \begin{cases} I, & z \in \mathcal{K}_{\text{exp}} \\ -I, & z \in \mathcal{K}^*_{\text{exp}} \\ \mathbf{diag}(1, I_+(z_2), I_+(z_3)), & z_1, z_2 < 0, \end{cases}$$

where $I_+(z_i),\ i = 2, 3$, is the indicator function of the nonnegative orthant, *i.e.*, it equals 1 if $z_i \geq 0$ and 0 otherwise. For the fourth case, the projection $P_{\mathcal{K}_{\text{exp}}}(z)$ is the solution to the optimization problem given in (12). Now observe that (i) the optimization problem (12) is, in fact, convex, since the constraint $\tilde{z}_2 > 0$ is really just implied by the domain of the function $\exp(\tilde{z}_1/\tilde{z}_2)$; (ii) the optimization problem (12) is feasible, since $z_1^\star = 1,\ z_2^\star = 1,\ z_3^\star = \exp(1)$ satisfies the constraint; and (iii) we can obtain a solution to the optimization problem (12), by using a Newton method (Parikh & Boyd, 2014, Section 6.3.4).

The rest of the proof relies on the KKT conditions for the optimization problem (12), as well as differentials (see, *e.g.*, Magnus & Neudecker (1995)). The Lagrangian of the optimization problem (12) is given by

$$(1/2)\|\tilde{z} - z\|_2^2 + \nu(\tilde{z}_2 \exp(\tilde{z}_1/\tilde{z}_2) - \tilde{z}_3),$$

where $\nu \in \mathbf{R}$ is the dual variable. Thus, we get that the KKT conditions for the optimization problem (12), at a solution $\gamma^\star = (z_1^\star, z_2^\star, z_3^\star, \nu^\star)$, are

$$z_1^\star - z_1 + \nu^\star \exp(z_1^\star/z_2^\star) = 0 \tag{S.2}$$
$$z_2^\star - z_2 + \nu^\star(\exp(z_1^\star/z_2^\star) - (z_1^\star/z_2^\star)\exp(z_1^\star/z_2^\star)) = 0 \tag{S.3}$$
$$z_3^\star - z_3 - \nu^\star = 0 \tag{S.4}$$
$$z_2^\star \exp(z_1^\star/z_2^\star) - z_3^\star = 0. \tag{S.5}$$

Now consider the differentials $dz_1^\star$, $dz_2^\star$, $dz_3^\star$, $dz_4^\star$ and dz_1, dz_2, dz_3, dz_4 of the KKT conditions (S.2) – (S.5); we get for the condition (S.2) that

$$\begin{aligned} & dz_1^\star - dz_1 + (d\nu^\star)\exp(z_1^\star/z_2^\star) + \nu^\star(d\exp(z_1^\star/z_2^\star)) = 0 \\ \iff\ & dz_1^\star - dz_1 + (d\nu^\star)\exp(z_1^\star/z_2^\star) + \nu^\star\exp(z_1^\star/z_2^\star)(d(z_1^\star/z_2^\star)) = 0 \\ \iff\ & dz_1^\star - dz_1 + (d\nu^\star)\exp(z_1^\star/z_2^\star) + \nu^\star\exp(z_1^\star/z_2^\star)\left(\frac{dz_1^\star}{z_2^\star} - \frac{z_1^\star(dz_2^\star)}{(z_2^\star)^2}\right) = 0 \\ \iff\ & \begin{bmatrix} 1+\frac{\nu^\star\exp(z_1^\star/z_2^\star)}{z_2^\star} & -\frac{\nu^\star\exp(z_1^\star/z_2^\star)z_1^\star}{(z_2^\star)^2} & 0 & \exp(z_1^\star/z_2^\star)\end{bmatrix}\begin{bmatrix} dz_1^\star \\ dz_2^\star \\ dz_3^\star \\ d\nu^\star\end{bmatrix} = dz_1. \end{aligned}$$

Repeating the above for the other conditions (S.3) – (S.5), we get that

$$\underbrace{\begin{bmatrix} 1+\frac{\nu^\star\exp(z_1^\star/z_2^\star)}{z_2^\star} & -\frac{\nu^\star\exp(z_1^\star/z_2^\star)z_1^\star}{(z_2^\star)^2} & 0 & \exp(z_1^\star/z_2^\star) \\ -\frac{\nu^\star\exp(z_1^\star/z_2^\star)z_1^\star}{(z_2^\star)^2} & 1+\frac{\nu^\star\exp(z_1^\star/z_2^\star)(z_1^\star)^2}{(z_2^\star)^3} & 0 & (1-z_1^\star/z_2^\star)\exp(z_1^\star/z_2^\star) \\ 0 & 0 & 1 & -1 \\ \exp(z_1^\star/z_2^\star) & (1-z_1^\star/z_2^\star)\exp(z_1^\star/z_2^\star) & -1 & 0 \end{bmatrix}}_{D}\underbrace{\begin{bmatrix} dz_1^\star \\ dz_2^\star \\ dz_3^\star \\ d\nu^\star\end{bmatrix}}_{d\gamma^\star} = \underbrace{\begin{bmatrix} dz_1 \\ dz_2 \\ dz_3 \\ d\nu\end{bmatrix}}_{d\gamma}, \tag{S.6}$$

i.e.,

$$D(d\gamma^\star) = d\gamma \quad \iff \quad d\gamma^\star = D^{-1}(d\gamma);$$

here, D is nonsingular, since the optimization problem (12) is feasible. So, by definition, the upper left 3x3 submatrix of D^{-1} is the Jacobian of the projection onto the exponential cone, for the fourth case. □

S.8. Intuition behind some of the regularity conditions for Theorem 4.1, Theorem 4.2, and Theorem 4.3

Here, we elaborate on a couple of the regularity assumptions stated in the main paper.

S.8.1. Regularity condition (A4)

Roughly speaking, the condition (A4) can be seen as requiring that the directional derivative of $\tilde{z} \mapsto \|F(\tilde{z})\|_2^2$ be bounded by $\alpha^{1/2}\|F(\tilde{z})\|_2^2$.

We list some (useful) functions satisfying (A4):

- The function $F(z) = z^2$, for $z \in \mathbf{R}$. To show that the function F satisfies (A4), we proceed by computing the required ingredients on both sides of (A4). Here, and for the rest of the section, we write $D_\Delta F^2(z)$ to mean the directional derivative of the function F squared, in the direction Δ, evaluated at z.

  We compute, for $z > 0$ and the Newton direction $\Delta = -1$, the left-hand side of (A4),

  $$D_\Delta F^2(z) = -4z^3,$$

  and the right-hand side of (A4),

  $$-\alpha^{1/2} 2z^3.$$

  So, satisfying (A4) means

  $$-4z^3 \le -\alpha^{1/2}2z^3 \iff 2 \ge \alpha^{1/2},$$

  which is certainly true. Repeating the argument for $z < 0$ and $\Delta = 1$ yields a similar result. (When $z = 0$, it is a solution.) Hence, $F(z) = z^2$ satisfies (A4).

- The function $F(z) = \max(z+1, cz+1)$, with $z \in \mathbf{R}$ and some $c > 0$.

  We have, for the left-hand side of (A4):
$$D_\Delta F^2(z=0) = -2c.$$

  We have, for the right-hand side of (A4):
$$\hat{F}(z=0, \Delta=-1) = J(z=0)\Delta = 1 \cdot (-1) = -1.$$

  So, satisfying (A4) means
$$-2c \leq -\alpha^{1/2} \iff c \geq \alpha^{1/2}/2.$$

  In words, functions that satisfy (A4) cannot have c too small.

- An argument similar the one used above for $F(z) = z^2$ can be used to show that the function $F(z) = |z|$ also satisfies (A4).

We also establish, by using the condition (A4), that the backtracking line search, used in Algorithm 1, terminates. Suppose, for contradiction, that the backtracking line search never terminates. Then, from the backtracking line search iteration described in Algorithm 1, we have, for all backtracking iterations k,
$$(\|F(z) + \gamma^{(k)}\Delta\|_2^2 - \|F(z)\|_2^2)/\gamma^{(k)} \geq -\alpha\|F(z)\|_2^2.$$

Taking the limit as $k \to \infty$, we get
$$D_\Delta\|F(z)\|_2^2 \geq -\alpha\|F(z)\|_2^2. \tag{S.7}$$

On the other hand, expanding the right-hand side of (A4) gives
$$\begin{aligned}
\alpha^{1/2}F(z)^T\hat{F}(z,\Delta) &= \alpha^{1/2}\left(F(z)^T(\hat{F}(z,\Delta) + F(z)) - F(z)^T F(z)\right) && \text{(S.8)}\\
&\leq \alpha^{1/2}\left(\|F(z)\|_2\|\hat{F}(z,\Delta) + F(z)\|_2 - \|F(z)\|_2^2\right) && \text{(S.9)}\\
&\leq \alpha^{1/2}\left(\|F(z)\|_2\varepsilon\|F(z)\|_2 - \|F(z)\|_2^2\right) && \text{(S.10)}\\
&\leq \alpha^{1/2}\left((1-\alpha^{1/2})\|F(z)\|_2^2 - \|F(z)\|_2^2\right) && \text{(S.11)}\\
&= -\alpha^{1/4}\|F(z)\|_2^2. && \text{(S.12)}
\end{aligned}$$

Putting (A4) and (S.12) above together immediately gives
$$D_\Delta\|F(z)\|_2^2 \leq -\alpha^{1/4}\|F(z)\|_2^2. \tag{S.13}$$

But putting (S.7) and (S.13) together gives
$$-\alpha\|F(z)\|_2^2 \leq D_\Delta\|F(z)\|_2^2 \leq -\alpha^{1/4}\|F(z)\|_2^2,$$

a contradiction, since $\alpha \in (0,1)$.

S.8.2. Regularity condition (A5)

Roughly speaking, the condition (A5) says that the Newton step on each iteration cannot be too large.

S.9. Proof of Theorem 4.1

Proof. We begin by recalling the condition under which backtracking line search continues, for a particular iteration of Newton's method; this happens as long as (see Algorithm 1)

$$\|F(z^{(i)} + t^{(i)}\Delta^{(i)})\|_2^2 \geq (1 - \alpha t^{(i)})\|F(z^{(i)})\|_2^2. \tag{S.14}$$

This means that when backtracking line search terminates, we get that

$$0 \leq \|F(z^{(i+1)})\|_2^2 < (1 - \alpha t^{(i)})\|F(z^{(i)})\|_2^2 < \|F(z^{(i)})\|_2^2. \tag{S.15}$$

(To be clear, in order to get the second inequality here, we used the fact that backtracking line search terminates after (S.14) in Algorithm 1 no longer holds.) In order to get the third inequality here, we used the simple fact that $0 < 1 - \alpha t^{(i)} \leq 1$, since $0 < \alpha < 1/2$ and $0 < t^{(i)} \leq 1$. So, we have shown that the sequence $(\|F(z^{(i)})\|_2^2)_{i=1}^\infty$ is both bounded below and decreasing. Note that this is just a sequence in **R**, and thus, by the monotone convergence theorem, it converges. Furthermore, since every convergent sequence in **R** is Cauchy, we get that

$$\lim_{i \to \infty} \left(\|F(z^{(i)})\|_2^2 - \|F(z^{(i+1)})\|_2^2 \right) = 0. \tag{S.16}$$

On the other hand, by rearranging the second inequality in (S.15), we get that

$$\|F(z^{(i)})\|_2^2 - \|F(z^{(i+1)})\|_2^2 > \alpha t^{(i)}\|F(z^{(i)})\|_2^2 \geq 0. \tag{S.17}$$

So, (S.16) along with taking the $\limsup_{i \to \infty}$ on both sides of (S.17) yields that $\lim_{i \to \infty} \alpha t^{(i)}\|F(z^{(i)})\|_2^2 = 0$. But assumption (A1) says that $\limsup_{i \to \infty} t^{(i)} \to t > 0$, and since $\alpha > 0$, we get that $\lim_{i \to \infty} \tilde{t}\|F(z^{(i)})\|_2^2 = 0$, for some $\tilde{t} > 0$, and so $\lim_{i \to \infty} \|F(z^{(i)})\|_2^2 = 0$, which implies that $\lim_{i \to \infty} F(z^{(i)}) = 0$, as claimed. □

S.10. Proof of Theorem 4.2

Proof. First of all, by the assumption that $(z^{(i)})_{i=1}^\infty$ is convergent and assumption (A5), we must have that

$$0 \leq \|\Delta^{(i)}\|_2 \leq \frac{1}{C_2}\|\hat{F}(z^{(i)}, \Delta^{(i)})\|_2 \leq \frac{\varepsilon + 1}{C_2}\|F(z^{(i)})\|_2, \tag{S.18}$$

where the second inequality here follows by rearranging (A5), and the third inequality follows from (20), as well as the triangle inequality: after computing $\Delta^{(i)}$ on Newton iteration i, we are assured that

$$\begin{aligned} &\|F(z^{(i)}) + \hat{F}(z^{(i)}, \Delta^{(i)})\|_2 \leq \varepsilon\|F(z^{(i)})\|_2 \\ \implies &\|\hat{F}(z^{(i)}, \Delta^{(i)})\|_2 - \|F(z^{(i)})\|_2 \leq \varepsilon\|F(z^{(i)})\|_2 \\ &\iff \|\hat{F}(z^{(i)}, \Delta^{(i)})\|_2 \leq (\varepsilon + 1)\|F(z^{(i)})\|_2. \end{aligned}$$

Hence, since

$$\sup_{j,\ell} \mathbf{dist}(\Delta^{(j)}, \Delta^{(\ell)}) \quad \leq \quad \sup_j \|\Delta^{(j)}\|_2 + \sup_\ell \|\Delta^{(\ell)}\|_2,$$

and because the right-hand side here is bounded (as per (S.18), as well as the fact that $(\|F(z^{(i)})\|_2^2)_{i=1}^\infty$ is decreasing), we can conclude that the sequence $(\Delta^{(i)})_{i=1}^\infty$ is bounded. (We used the Euclidean distance here.)

By the Bolzano-Weierstrass theorem (for Euclidean spaces), this sequence contains a convergent subsequence; let $(\Delta^{(i)})_{i \in \mathcal{S}}$, for some countable set $\mathcal{S}$, be this subsequence. Define $\gamma^{(i)} = t^{(i)}/\beta$, *i.e.*, $\gamma^{(i)}$ is the last $t^{(i)}$ for which (S.14) was actually true (*i.e.*, when checked at the start of the $(i+1)$th Newton iteration). Then we get

$$\|F(z^{(i)} + \gamma^{(i)}\Delta^{(i)})\|_2^2 - \|F(z^{(i)})\|_2^2 \geq -\alpha\gamma^{(i)}\|F(z^{(i)})\|_2^2;$$

dividing through by $\gamma^{(i)}$ and taking limits gives (observe that, from assumption (A2), $\limsup_{i \to \infty} t^{(i)} = 0 \implies \lim_{i \to \infty} t^{(i)} = 0$)

$$\begin{aligned} -\alpha\|F(z)\|_2^2 &\leq \lim_{i,j \to \infty,\, j \in \mathcal{S}} \frac{\|F(z^{(i)} + \gamma^{(i)}\Delta^{(j)})\|_2^2 - \|F(z^{(i)})\|_2^2}{\gamma^{(i)}} && \text{(S.19)} \\ &\leq \lim_{i,j \to \infty,\, j \in \mathcal{S}} \alpha^{1/2} F(z^{(i)})^T \hat{F}(z^{(i)}, \Delta^{(j)}), && \text{(S.20)} \end{aligned}$$

with the second line here following by assumption (A4). Expanding the right-hand side of (S.20), we get

$$\begin{aligned}\alpha^{1/2} F(z^{(i)})^T \hat{F}(z^{(i)}, \Delta^{(j)}) &= \alpha^{1/2} F(z^{(i)})^T \left(\hat{F}(z^{(i)}, \Delta^{(j)}) + F(z^{(i)})\right) - \alpha^{1/2} F(z^{(i)})^T F(z^{(i)}) \\ &\leq \alpha^{1/2} \|F(z^{(i)})\|_2 \|F(z^{(i)}) + \hat{F}(z^{(i)}, \Delta^{(j)})\|_2 - \alpha^{1/2} \|F(z^{(i)})\|_2^2 \\ &\leq \alpha^{1/2} \varepsilon \|F(z^{(i)})\|_2^2 - \alpha^{1/2} \|F(z^{(i)})\|_2^2 \\ &= -\alpha^{1/2} \|F(z^{(i)})\|_2^2 (1 - \varepsilon),\end{aligned}$$

with the second line following from the Cauchy-Schwarz inequality, and the third from (20). So, we obtain for the right-hand side of (S.20) that

$$\lim_{i,j \to \infty,\ j \in \mathcal{S}} \alpha^{1/2} F(z^{(i)})^T \hat{F}(z^{(i)}, \Delta^{(j)}) \leq -\alpha^{1/2} \|F(z)\|_2^2 (1 - \varepsilon). \tag{S.21}$$

Putting together (S.19) and (S.21), we get that

$$-\alpha \|F(z)\|_2^2 \leq -\alpha^{1/2} \|F(z)\|_2^2 (1 - \varepsilon) \quad \iff \quad 0 \geq \alpha^{1/2} \|F(z)\|_2^2 \left((1 - \varepsilon) - \alpha^{1/2}\right).$$

Now, by assumption (A3), we require that $\varepsilon < 1 - \alpha^{1/2} \iff (1 - \varepsilon) - \alpha^{1/2} > 0$; thus, we must have that $\|F(z)\|_2^2 = 0 \iff F(z) = 0$, as claimed. □

S.11. Proof of Theorem 4.3

Proof. The theorem establishes that the iterates $(z^{(i)})_{i=1}^{\infty}$ generated by Algorithm 1 are *locally quadratically convergent*, *i.e.*, we get, for large enough i and some $C > 0$, that

$$\lim_{i \to \infty} \frac{|z^{(i+1)} - z|}{(z^{(i)} - z)^2} = C.$$

Let $\mathbf{res}(i) = F(z^{(i)}) + J(z^{(i)})\Delta^{(i)}$, for convenience. We begin by making two useful observations.

First, using the second part of assumption (A6), we get that

$$\begin{aligned}\|F(z^{(i)}) - \mathbf{res}(i)\|_2 &= \|J(z^{(i)})\Delta^{(i)}\|_2 \\ &\leq \|J(z^{(i)})\|_2 \|\Delta^{(i)}\|_2 \\ &\leq C_3 \|\Delta^{(i)}\|_2.\end{aligned} \tag{S.22}$$

On the other hand, using the triangle inequality as well as (20), we get that

$$\begin{aligned}\|F(z^{(i)}) - \mathbf{res}(i)\|_2 &\geq \|F(z^{(i)})\|_2 - \|\mathbf{res}(i)\|_2 \\ &\geq \|F(z^{(i)})\|_2 - \varepsilon \|F(z^{(i)})\|_2 \\ &\geq (1 - \varepsilon) \|F(z^{(i)})\|_2.\end{aligned} \tag{S.23}$$

So, putting together (S.22) and (S.23), we get that

$$(1 - \varepsilon) \|F(z^{(i)})\|_2 \leq C_3 \|\Delta^{(i)}\|_2 \quad \implies \quad \|F(z^{(i)})\|_2 \leq C_4 \|\Delta^{(i)}\|_2,$$

for some constant $C_4 > 0$, since $1 - \varepsilon > 0$. Squaring both sides, it follows that

$$\begin{aligned}&\|F(z^{(i)})\|_2 \leq C_4 \|\Delta^{(i)}\|_2 \\ \implies\ &\|F(z^{(i)})\|_2^2 \leq C_4^2 \|\Delta^{(i)}\|_2^2 \\ \implies\ &\|\mathbf{res}(i)\|_2 \leq C_5 \|\Delta^{(i)}\|_2^2 \\ \implies\ &\frac{\|\mathbf{res}(i)\|_2}{\|\Delta^{(i)}\|_2^2} \leq C_5,\end{aligned} \tag{S.24}$$

where $C_5 > 0$ is some constant, and the third line follows because (20) and assumption (A3) tell us that $\|\mathbf{res}(i)\|_2 \leq C\|F(z^{(i)})\|_2^2$ for some constant $C > 0$. Finally, Facchinei & Kanzow (1997, Theorem 2.5) and the second part of assumption (A6) tell us that the sequence of iterates $(z^{(i)})_{i=1}^{\infty} \to z$ converges quadratically, with $F(z) = 0$, as claimed. □

S.12. Further details on the minimum variance portfolio optimization example

Here, we elaborate on putting the minimum variance portfolio optimization problem (22) into the cone form of (1).

First, we rewrite the minimum variance portfolio optimization problem (22) as

$$\begin{array}{ll} \underset{\theta \in \mathbf{R}^p,\, w \in \mathbf{R}}{\text{minimize}} & w \\ \text{subject to} & \left\| \begin{bmatrix} 2\Sigma^{1/2}\theta \\ 1 - w \end{bmatrix} \right\|_2 \leq 1 + w \\ & 1 \leq \mathbf{1}^T\theta \leq 1, \end{array}$$

where we used the simple fact (Lobo et al., 1998, Equation 8) that

$$\alpha^T\alpha \leq \gamma\delta \quad \iff \quad \left\| \begin{bmatrix} 2\alpha \\ \gamma - \delta \end{bmatrix} \right\|_2 \leq \gamma + \delta,$$

for some vector α and nonnegative constants θ, γ (for us, $\alpha = \Sigma^{1/2}\theta$, $\gamma = 1$, and $\delta = w$). Then, we rewrite the above problem as

$$\begin{array}{ll} \underset{x \in \mathbf{R}^{p+1}}{\text{minimize}} & c^T x \\ \text{subject to} & \|G_1 x + h\|_2 \leq q^T x + z \\ & G_2 x \leq 1, \quad G_3 x \leq -1, \end{array}$$

where we defined

$$\begin{aligned} x &= \begin{bmatrix} \theta \\ w \end{bmatrix} \\ c &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ G_1 &= \begin{bmatrix} 2\Sigma^{1/2} & 0 \\ 0 & -1 \end{bmatrix}, \quad h = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ q &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad z = 1 \\ G_2 &= \begin{bmatrix} \mathbf{1}^T & 0 \end{bmatrix}, \quad G_3 = \begin{bmatrix} -\mathbf{1}^T & 0 \end{bmatrix}. \end{aligned}$$

Finally, we just use

$$A = \begin{bmatrix} -G_1 \\ -q^T \\ G_2 \\ G_3 \end{bmatrix}, \quad b = \begin{bmatrix} h \\ z \\ 1 \\ -1 \end{bmatrix}, \quad \mathcal{K} = \mathcal{K}_{\text{soc}}^{p+2} \times \mathcal{K}_{\text{no}} \times \mathcal{K}_{\text{no}},$$

to get the cone form of (1); here, $\mathcal{K}_{\text{soc}}^{p+2}$ denotes the $(p+1)$-dimensional second-order cone.

S.13. Further details on the ℓ_1-penalized logistic regression example

Here, we elaborate on putting the ℓ_1-penalized logistic regression problem (23) into the cone form of (1). To keep the notation light, we write

$$z_i = y_i X_{i\cdot}\theta.$$

Now, for $i = 1, \ldots, N$, we use the simple fact (Serrano, 2015, Section 9.4.1) that

$$\log\left(\sum_i \exp(\alpha_i)\right) \leq -\theta \quad \iff \quad \sum_i \exp(\alpha_i + \theta) \leq 1,$$

for $\alpha_i, \theta \in \mathbf{R}$, in order to conclude that

$$\log(\exp(0) + \exp(z_i)) \le w_i \quad \iff \quad \exp(-w_i) + \exp(z_i - w_i) \le 1, \tag{S.25}$$

where the $w_i \in \mathbf{R}$ are some variables that we will introduce, later on. Next, we "split" the right-hand side of (S.25) into the following set of constraints:

$$\begin{aligned}
\exp(-w_i) \le \ell_i \quad &\iff \quad \begin{bmatrix} -w_i \\ 1 \\ \ell_i \end{bmatrix} \in \mathcal{K}_{\text{exp}}, \quad i = 1, \dots, N, \\
\exp(z_i - w_i) \le q_i \quad &\iff \quad \begin{bmatrix} z_i - w_i \\ 1 \\ q_i \end{bmatrix} \in \mathcal{K}_{\text{exp}}, \quad i = 1, \dots, N, \\
\ell_i + q_i \le 1, \quad i = 1, \dots, N,
\end{aligned}$$

where $\ell_i, q_i \in \mathbf{R}$ are more new variables. Thus, we can write the ℓ_1-penalized logistic regression problem (23) as

$$\begin{array}{ll}
\underset{\substack{\theta \in \mathbf{R}^p,\, w \in \mathbf{R}^N, \\ t \in \mathbf{R}^p,\, \ell \in \mathbf{R}^N, \\ q \in \mathbf{R}^N}}{\text{minimize}} & \mathbf{1}^T w + \lambda \mathbf{1}^T t \\
\text{subject to} & \begin{bmatrix} -w_i \\ 1 \\ \ell_i \end{bmatrix} \in \mathcal{K}_{\text{exp}}, \quad i = 1, \dots, N \\
& \begin{bmatrix} y_i X_{i\cdot} \theta - w_i \\ 1 \\ q_i \end{bmatrix} \in \mathcal{K}_{\text{exp}}, \quad i = 1, \dots, N \\
& \ell + q \le \mathbf{1} \\
& -t \le \theta \le t.
\end{array}$$

Finally, to get the cone form of (1), we use

$$x = \begin{bmatrix} \theta \\ w \\ t \\ \ell \\ q \end{bmatrix},$$

$$c = \begin{bmatrix} 0 \\ \mathbf{1} \\ \lambda\mathbf{1} \\ 0 \\ 0 \end{bmatrix},$$

$$A = \begin{bmatrix} & & G_1 & & \\ & & \vdots & & \\ & & G_N & & \\ & & H_1 & & \\ & & \vdots & & \\ & & H_N & & \\ 0 & 0 & 0 & I & I \\ -I & 0 & -I & 0 & 0 \\ I & 0 & -I & 0 & 0 \end{bmatrix},$$

$$G_i = \begin{bmatrix} 0 & e_i^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -e_i^T & 0 \end{bmatrix}, \quad H_i = \begin{bmatrix} -y_i X_{i\cdot} & e_i^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -e_i^T \end{bmatrix}, \quad i = 1, \ldots, N,$$

$$b = \begin{bmatrix} h \\ \vdots \\ h \\ h \\ \vdots \\ h \\ \mathbf{1} \\ 0 \\ 0 \end{bmatrix}, \quad h = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},$$

$$\mathcal{K} = \underbrace{\mathcal{K}_{\text{exp}} \times \cdots \times \mathcal{K}_{\text{exp}}}_{N} \times \underbrace{\mathcal{K}_{\text{exp}} \times \cdots \times \mathcal{K}_{\text{exp}}}_{N} \times \mathcal{K}_{\text{no}}^N \times \mathcal{K}_{\text{no}}^p \times \mathcal{K}_{\text{no}}^p;$$

here, $e_i,\ i = 1, \ldots, N$ denotes the ith standard basis vector in $\mathbf{R}^N$, and $\mathcal{K}_{\text{no}}^i$ denotes the i-dimensional nonnegative orthant.

S.14. Further details on the robust PCA example

Here, we elaborate on putting the robust PCA problem (24) into the cone form of (1).

First, we observe that, using duality arguments (see, *e.g.*, Fazel et al. (2001, Section 3) or Recht et al. (2010, Proposition 2.1)), we can rewrite the robust PCA problem (24) as

$$\begin{array}{ll} \underset{\substack{W_1 \in \mathbf{R}^{N \times N}, W_2 \in \mathbf{R}^{p \times p}, \\ t \in \mathbf{R}^{Np},\, L, S \in \mathbf{R}^{N \times p}}}{\text{minimize}} & (1/2)(\mathbf{tr}(W_1) + \mathbf{tr}(W_2)) \\ \text{subject to} & -t \le \mathbf{vec}(S) \le t \\ & \mathbf{1}^T t \le \lambda \\ & L + S = X \\ & \begin{bmatrix} W_1 & L \\ L^T & W_2 \end{bmatrix} \succeq 0. \end{array}$$

To get the cone form of (1), we use

$$
\begin{aligned}
x &= \begin{bmatrix} \mathbf{vec}(W_1) \\ \mathbf{vec}(W_2) \\ t \\ \mathbf{vec}(L) \\ \mathbf{vec}(S) \end{bmatrix}, \\
c &= \begin{bmatrix} (1/2)\,\mathbf{vec}(I) \\ (1/2)\,\mathbf{vec}(I) \\ 0 \\ 0 \\ 0 \end{bmatrix}, \\
A &= \begin{bmatrix} 0 & 0 & -I & 0 & -I \\ 0 & 0 & -I & 0 & I \\ 0 & 0 & \mathbf{1}^T & 0 & 0 \\ 0 & 0 & 0 & I & I \\ 0 & 0 & 0 & -I & -I \\ G_{W_1} & G_{W_2} & 0 & G_L & 0 \end{bmatrix}, \\
G_{W_1} &= \begin{bmatrix} \mathbf{vec}(G_{W_1}^{(1,1)}) & \mathbf{vec}(G_{W_1}^{(2,1)}) & \cdots & \mathbf{vec}(G_{W_1}^{(N-1,N)}) & \mathbf{vec}(G_{W_1}^{(N,N)}) \end{bmatrix}, \\
&\quad \text{where } G_{W_1}^{(i,j)} \text{ is 0 except with the } (i,j)\text{th entry of its upper left } N \times N \text{ block set to 1}, \\
G_{W_2} &= \begin{bmatrix} \mathbf{vec}(G_{W_2}^{(1,1)}) & \mathbf{vec}(G_{W_2}^{(2,1)}) & \cdots & \mathbf{vec}(G_{W_2}^{(p-1,p)}) & \mathbf{vec}(G_{W_2}^{(p,p)}) \end{bmatrix}, \\
&\quad \text{where } G_{W_2}^{(i,j)} \text{ is 0 except with the } (i,j)\text{th entry of its bottom right } p \times p \text{ block set to 1}, \\
G_L &= \begin{bmatrix} \mathbf{vec}(G_L^{(1,1)}) & \mathbf{vec}(G_L^{(2,1)}) & \cdots & \mathbf{vec}(G_L^{(N-1,p)}) & \mathbf{vec}(G_L^{(N,p)}) \end{bmatrix}, \\
&\quad \text{where } G_L^{(i,j)} \text{ is 0 except with the } (i,j)\text{th entry of its upper right } N \times p \text{ block} \\
&\quad \text{and the } (j,i)\text{th entry of its lower left } p \times N \text{ block set to 1}, \\
b &= \begin{bmatrix} 0 \\ 0 \\ \lambda \\ \mathbf{vec}(X) \\ -\mathbf{vec}(X) \\ 0 \end{bmatrix}, \\
\mathcal{K} &= \mathcal{K}_{\text{no}}^{Np} \times \mathcal{K}_{\text{no}}^{Np} \times \mathcal{K}_{\text{no}} \times \mathcal{K}_{\text{no}}^{Np} \times \mathcal{K}_{\text{no}}^{Np} \times \mathcal{K}_{\text{psd}}^{N+p}.
\end{aligned}
$$

Here, $\mathcal{K}_{\text{psd}}^i$ denotes the $(i \times i)$-dimensional positive semidefinite cone. Also, observe that the last row of A, b above encodes the constraint

$$
\begin{bmatrix} W_1 & L \\ L^T & W_2 \end{bmatrix} \in \mathcal{K}_{\text{psd}}^{N+p},
$$

which we can write as a linear matrix inequality (Andersen et al., 2011, Equation 1.7):

$$
\begin{aligned}
\begin{bmatrix} W_1 & L \\ L^T & W_2 \end{bmatrix} \succeq 0 &\iff \sum_{i,j} G_{W_1}^{(i,j)} (W_1)_{ij} + \sum_{i,j} G_{W_2}^{(i,j)} (W_2)_{ij} + \sum_{i,j} G_L^{(i,j)} L_{ij} \succeq 0 \\
&\iff \begin{bmatrix} G_{W_1} & G_{W_2} & 0 & G_L & 0 \end{bmatrix} x \succeq 0.
\end{aligned}
$$

Expression for the matrix D^{-1}, used in the proof of Lemma 3.2:

$$D^{-1} = (1/k) \cdot \begin{bmatrix} -1 - \frac{e^{z_1^\star/z_2^\star}\nu^\star (z_1^\star)^2}{(z_2^\star)^3} - \frac{e^{z_1^\star/z_2^\star} e^{(z_1^\star)^3}}{(z_2^\star)^3} + \frac{2e^{z_1^\star/z_2^\star} e^{(z_1^\star)^2}}{(z_2^\star)^2} - \frac{e^{z_1^\star/z_2^\star} e^{z_1^\star}}{z_2^\star} & -\frac{e^{z_1^\star/z_2^\star}\nu^\star z_1^\star}{(z_2^\star)^2} - \frac{e^{z_1^\star/z_2^\star} e^{(z_1^\star)^2}}{(z_2^\star)^2} + \frac{e^{z_1^\star/z_2^\star} e^{z_1^\star}}{z_2^\star} & -e^{z_1^\star/z_2^\star} - \frac{e^{\frac{2z_1^\star}{z_2^\star}}\nu^\star z_1^\star}{(z_2^\star)^2} & -e^{z_1^\star/z_2^\star} - \frac{e^{\frac{2z_1^\star}{z_2^\star}}\nu^\star z_1^\star}{(z_2^\star)^2} \\ e^{\frac{2z_1^\star}{z_2^\star}} - \frac{e^{z_1^\star/z_2^\star}\nu^\star z_1^\star}{(z_2^\star)^2} - \frac{e^{\frac{2z_1^\star}{z_2^\star}} z_1^\star}{z_2^\star} & -1 - e^{\frac{2z_1^\star}{z_2^\star}} - \frac{e^{z_1^\star/z_2^\star}\nu^\star}{z_2^\star} & -e^{z_1^\star/z_2^\star} - \frac{e^{\frac{2z_1^\star}{z_2^\star}}\nu^\star}{z_2^\star} + \frac{e^{z_1^\star/z_2^\star} z_1^\star}{z_2^\star} & -e^{z_1^\star/z_2^\star} - \frac{e^{\frac{2z_1^\star}{z_2^\star}}\nu^\star}{z_2^\star} + \frac{e^{z_1^\star/z_2^\star} z_1^\star}{z_2^\star} \\ -e^{z_1^\star/z_2^\star} + \frac{e^{z_1^\star/z_2^\star} e^{\nu^\star (z_1^\star)^3}}{(z_2^\star)^4} - \frac{e^{\frac{2z_1^\star}{z_2^\star}}\nu^\star (z_1^\star)^2}{(z_2^\star)^3} - \frac{e^{z_1^\star/z_2^\star} e^{\nu^\star (z_1^\star)^2}}{(z_2^\star)^3} & \frac{e^{z_1^\star/z_2^\star} e^{\nu^\star (z_1^\star)^2}}{(z_2^\star)^3} - \frac{e^{\frac{2z_1^\star}{z_2^\star}}\nu^\star z_1^\star}{(z_2^\star)^2} - \frac{e^{z_1^\star/z_2^\star} e^{\nu^\star z_1^\star}}{(z_2^\star)^2} + \frac{e^{(z_1^\star)^2}}{(z_2^\star)^2} - \frac{e^{z_1^\star}}{z_2^\star} & -e^{\frac{2z_1^\star}{z_2^\star}} + \frac{e^{\frac{2z_1^\star}{z_2^\star}} e^{\nu^\star (z_1^\star)^2}}{(z_2^\star)^3} - \frac{e^{z_1^\star/z_2^\star} e^{(z_1^\star)^3}}{(z_2^\star)^3} - \frac{e^{\frac{3z_1^\star}{z_2^\star}}\nu^\star z_1^\star}{(z_2^\star)^2} - \frac{e^{\frac{2z_1^\star}{z_2^\star}} e^{\nu^\star z_1^\star}}{(z_2^\star)^2} + \frac{2e^{z_1^\star/z_2^\star} e^{(z_1^\star)^2}}{(z_2^\star)^2} - \frac{e^{z_1^\star/z_2^\star} e^{z_1^\star}}{z_2^\star} & 1 + \frac{e^{z_1^\star/z_2^\star}\nu^\star (z_1^\star)^2}{(z_2^\star)^3} + \frac{e^{z_1^\star/z_2^\star}\nu^\star}{z_2^\star} \\ -e^{z_1^\star/z_2^\star} + \frac{e^{z_1^\star/z_2^\star} e^{\nu^\star (z_1^\star)^3}}{(z_2^\star)^4} - \frac{e^{\frac{2z_1^\star}{z_2^\star}}\nu^\star (z_1^\star)^2}{(z_2^\star)^3} - \frac{e^{z_1^\star/z_2^\star} e^{\nu^\star (z_1^\star)^2}}{(z_2^\star)^3} & \frac{e^{z_1^\star/z_2^\star} e^{\nu^\star (z_1^\star)^2}}{(z_2^\star)^3} - \frac{e^{\frac{2z_1^\star}{z_2^\star}}\nu^\star z_1^\star}{(z_2^\star)^2} - \frac{e^{z_1^\star/z_2^\star} e^{\nu^\star z_1^\star}}{(z_2^\star)^2} + \frac{e^{(z_1^\star)^2}}{(z_2^\star)^2} - \frac{e^{z_1^\star}}{z_2^\star} & 1 + \frac{e^{z_1^\star/z_2^\star}\nu^\star (z_1^\star)^2}{(z_2^\star)^3} + \frac{e^{z_1^\star/z_2^\star}\nu^\star}{z_2^\star} & 1 + \frac{e^{z_1^\star/z_2^\star}\nu^\star (z_1^\star)^2}{(z_2^\star)^3} + \frac{e^{z_1^\star/z_2^\star}\nu^\star}{z_2^\star} \end{bmatrix}, \tag{S.26}$$

$$\text{where } k = -1 - e^{\frac{2z_1^\star}{\hat{z}_2}} - \frac{e^{z_1^\star/z_2^\star}\nu^\star (z_1^\star)^2}{(z_2^\star)^3} + \frac{e^{\frac{2z_1^\star}{z_2^\star}} e^{\nu^\star (z_1^\star)^2}}{(z_2^\star)^3} - \frac{e^{z_1^\star/z_2^\star} e^{(z_1^\star)^3}}{(z_2^\star)^3} - \frac{e^{\frac{3z_1^\star}{z_2^\star}}\nu^\star z_1^\star}{(z_2^\star)^2} - \frac{e^{\frac{2z_1^\star}{z_2^\star}} e^{\nu^\star z_1^\star}}{(z_2^\star)^2} + \frac{2e^{z_1^\star/z_2^\star} e^{(z_1^\star)^2}}{(z_2^\star)^2} - \frac{e^{z_1^\star/z_2^\star}\nu^\star}{z_2^\star} - \frac{e^{z_1^\star/z_2^\star} e^{z_1^\star}}{z_2^\star}.$$

References

Andersen, Martin, Dahl, Joachim, Liu, Zhang, and Vandenberghe, Lieven. Interior point methods for large-scale cone programming. *Optimization for machine learning*, pp. 55–83, 2011.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

Clarke, Frank. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

Facchinei, Francisco and Kanzow, Christian. A nonsmooth inexact Newton method for the solution of large-scale nonlinear complementarity problems. *Mathematical Programming*, 76(3):493–512, 1997.

Facchinei, Francisco and Pang, Jong-Shi. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2007.

Fazel, Maryam, Hindi, Haitham, and Boyd, Stephen. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pp. 4734–4739. IEEE, 2001.

Izmailov, Alexey and Solodov, Mikhail. *Newton-Type Methods for Optimization and Variational Problems*. Springer, 2014.

Lobo, Miguel Sousa, Vandenberghe, Lieven, Boyd, Stephen, and Lebret, Hervé. Applications of second-order cone programming. *Linear algebra and its applications*, 284(1):193–228, 1998.

Magnus, Jan and Neudecker, Heinz. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 1995.

O'Donoghue, Brendan, Chu, Eric, Parikh, Neal, and Boyd, Stephen. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, pp. 1–27, 2016.

Parikh, Neal and Boyd, Stephen. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

Qi, Liqun and Sun, Jie. A nonsmooth version of Newton's method. *Mathematical Programming*, 58(1-3):353–367, 1993.

Recht, Benjamin, Fazel, Maryam, and Parrilo, Pablo A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Serrano, Santiago. *Algorithms for Unsymmetric Cone Optimization and an Implementation for Problems with the Exponential Cone*. PhD thesis, Stanford University, 2015.