

---

# Appendix: Constant Nullspace Strong Convexity and Fast Convergence of Proximal Methods under High-Dimensional Settings

---

**Ian E.H. Yen    Cho-Jui Hsieh    Pradeep Ravikumar    Inderjit Dhillon**  
 Department of Computer Science  
 University of Texas at Austin  
 {ianyenyen, cjhsieh, pradeepr, inderjit}@cs.utexas.edu

## 1 Proof for properties of proximal operations

The proximal operator  $\mathbf{prox}(\cdot)$  is defined as

$$\mathbf{x}_{t+1} = \mathbf{prox}(\mathbf{x}_{t+\frac{1}{2}}) = \arg \min_{\mathbf{x}} h(\mathbf{x}) + \frac{M}{2} \|\mathbf{x} - \mathbf{x}_{t+\frac{1}{2}}\|_2^2. \quad (1)$$

**Lemma 1.** Define  $\Delta^P \mathbf{x} = \mathbf{x} - \mathbf{prox}(\mathbf{x})$ , the following properties hold for the proximal operation (1).

1.  $M\Delta^P \mathbf{x} \in \partial h(\mathbf{prox}(\mathbf{x}))$ .
2.  $\|\mathbf{prox}(\mathbf{x}_1) - \mathbf{prox}(\mathbf{x}_2)\|_2^2 \leq \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 - \|\Delta^P \mathbf{x}_1 - \Delta^P \mathbf{x}_2\|_2^2$ .

*Proof.* The first property follows directly from the optimality condition of (1). The second property holds since for  $M\Delta^P \mathbf{x}_1 \in \partial h(\mathbf{prox}(\mathbf{x}_1))$ ,  $M\Delta^P \mathbf{x}_2 \in \partial h(\mathbf{prox}(\mathbf{x}_2))$  we have  $\langle M\Delta^P \mathbf{x}_1 - M\Delta^P \mathbf{x}_2, \mathbf{prox}(\mathbf{x}_1) - \mathbf{prox}(\mathbf{x}_2) \rangle \geq 0$ , and thus,

$$\begin{aligned} \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 &= \|(\mathbf{prox}(\mathbf{x}_1) - \mathbf{prox}(\mathbf{x}_2)) + (\Delta^P \mathbf{x}_1 - \Delta^P \mathbf{x}_2)\|_2^2 \\ &\geq \|\mathbf{prox}(\mathbf{x}_1) - \mathbf{prox}(\mathbf{x}_2)\|_2^2 + \|\Delta^P \mathbf{x}_1 - \Delta^P \mathbf{x}_2\|_2^2, \end{aligned}$$

which gives the second property. □

The proximal operator  $\mathbf{prox}_H(\cdot)$  is defined for any PSD matrix  $H$  as

$$\mathbf{prox}_H(\mathbf{x}) = \arg \min_{\mathbf{v}} h(\mathbf{v}) + \frac{1}{2} \|\mathbf{v} - \mathbf{x}\|_H^2. \quad (2)$$

**Lemma 2.** Define  $\Delta^P \mathbf{x} = \mathbf{x} - \mathbf{prox}_H(\mathbf{x})$ , the following properties hold for the proximal operator:

1.  $H\Delta^P \mathbf{x} \in \partial h(\mathbf{prox}_H(\mathbf{x}))$ .
2.  $\|\mathbf{prox}_H(\mathbf{x}_1) - \mathbf{prox}_H(\mathbf{x}_2)\|_H^2 \leq \|\mathbf{x}_1 - \mathbf{x}_2\|_H^2$ .

*Proof.* The first property follows directly from the optimality condition of (2). The second property holds since for  $H\Delta^P \mathbf{x}_1 \in \partial h(\mathbf{prox}_H(\mathbf{x}_1))$ ,  $H\Delta^P \mathbf{x}_2 \in \partial h(\mathbf{prox}_H(\mathbf{x}_2))$  we have  $\langle H\Delta^P \mathbf{x}_1 - H\Delta^P \mathbf{x}_2, \mathbf{prox}_H(\mathbf{x}_1) - \mathbf{prox}_H(\mathbf{x}_2) \rangle \geq 0$ , and thus,

$$\begin{aligned} \|\mathbf{x}_1 - \mathbf{x}_2\|_H^2 &= \|(\mathbf{prox}_H(\mathbf{x}_1) - \mathbf{prox}_H(\mathbf{x}_2)) + (\Delta^P \mathbf{x}_1 - \Delta^P \mathbf{x}_2)\|_H^2 \\ &\geq \|\mathbf{prox}_H(\mathbf{x}_1) - \mathbf{prox}_H(\mathbf{x}_2)\|_H^2 + \|\Delta^P \mathbf{x}_1 - \Delta^P \mathbf{x}_2\|_H^2 \\ &\geq \|\mathbf{prox}_H(\mathbf{x}_1) - \mathbf{prox}_H(\mathbf{x}_2)\|_H^2, \end{aligned}$$

where the second inequality follows from the PSD of  $H$ . □

## 2 Proof of Lemma 3

**Lemma 3** (Optimal Set). *Let  $\bar{\mathcal{E}}$  be the active set at optimal and  $\bar{\mathcal{E}}^+ = \{j \mid \|\Pi_{\mathcal{M}_j}(\bar{\rho})\|_* = \lambda\}$  be its augmented set (which is unique since  $\bar{\rho}$  is unique) such that  $\Pi_{\mathcal{M}_j}(\bar{\rho}) = \lambda \bar{\mathbf{a}}_j$ ,  $j \in \bar{\mathcal{E}}^+$ . The optimal solutions then form a polyhedral set*

$$\bar{\mathcal{X}} = \{\mathbf{x} \mid \Pi_{\mathcal{T}}(\mathbf{x}) = \bar{\mathbf{z}} \text{ and } \mathbf{x} \in \bar{\mathcal{O}}\}, \quad (3)$$

where  $\bar{\mathcal{O}} = \{\mathbf{x} \mid \mathbf{x} = \sum_{j \in \bar{\mathcal{E}}^+} c_j \bar{\mathbf{a}}_j, c_j \geq 0, j \in \bar{\mathcal{E}}^+\}$  is the set of  $\mathbf{x}$  with  $\bar{\rho} \in \partial h(\mathbf{x})$ .

*Proof.* The optimality condition are  $\mathbf{g}(\mathbf{x}) = \bar{\mathbf{g}}$  and  $\bar{\rho} \in \partial h(\mathbf{x})$  by Theorem 1. Since  $\Pi_{\mathcal{T}}(\mathbf{x}) = \bar{\mathbf{z}}$ , we have  $\mathbf{g}(\mathbf{x}) = \bar{\mathbf{g}}$  already. Therefore, we only need to show that  $\bar{\rho} \in \partial h(\mathbf{x})$  iff  $\mathbf{x} \in \bar{\mathcal{O}}$ .

Suppose  $\bar{\rho} \in \partial h(\mathbf{x})$ . Then for  $j \notin \bar{\mathcal{E}}^+$ , we know  $\|\Pi_{\mathcal{M}_j}(\bar{\rho})\|_* < 1$ , which means  $\Pi_{\mathcal{M}_j}(\mathbf{x}) = 0$ , and for  $j \in \bar{\mathcal{E}}^+$ , we know  $\Pi_{\mathcal{M}_j}(\bar{\rho}) = \lambda \bar{\mathbf{a}}_j$ , which means  $\Pi_{\mathcal{M}_j}(\mathbf{x})$  can be  $\mathbf{0}$  or  $c_j \bar{\mathbf{a}}_j$  for some  $c_j > 0$ . Therefore,  $\mathbf{x}$  must have the form  $\mathbf{x} = \sum_{j \in \bar{\mathcal{E}}^+} c_j \bar{\mathbf{a}}_j, c_j \geq 0, j \in \bar{\mathcal{E}}^+$ .

Now for the other direction, suppose  $\mathbf{x} = \sum_{j \in \bar{\mathcal{E}}^+} c_j \bar{\mathbf{a}}_j, c_j \geq 0, j \in \bar{\mathcal{E}}^+$  and  $\mathcal{E} \subseteq \bar{\mathcal{E}}^+$  is the set for which  $c_j > 0, j \in \mathcal{E}$ . Then since  $\|\Pi_{\mathcal{M}_j}(\bar{\rho})\|_* \leq 1, j \notin \mathcal{E}$  and for  $j \in \mathcal{E} \subseteq \bar{\mathcal{E}}^+$  we have  $\Pi_{\mathcal{M}_j}(\bar{\rho}) = \lambda \bar{\mathbf{a}}_j$ , we conclude that  $\bar{\rho} \in \partial h(\mathbf{x})$ .  $\square$

## 3 Proof of Lemma 5

**Lemma 5.** *Let  $\bar{\mathcal{A}} = \text{span}(\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2, \dots, \bar{\mathbf{a}}_{|\bar{\mathcal{E}}^+|})$ . Suppose  $\|\mathbf{x}\| \leq R$  and  $\Pi_{\mathcal{M}_j}(\mathbf{x}) = \mathbf{0}$  for  $j \notin \bar{\mathcal{E}}^+$ . Then*

$$\lambda^2 \|\mathbf{x} - \Pi_{\bar{\mathcal{A}}}(\mathbf{x})\|_2^2 \leq R^2 \|\boldsymbol{\rho} - \bar{\boldsymbol{\rho}}\|_2^2,$$

where  $\boldsymbol{\rho} \in \partial h(\mathbf{x})$  and  $\bar{\boldsymbol{\rho}}$  is as defined in Theorem 1.

*Proof.* Since  $\Pi_{\mathcal{M}_j}(\mathbf{x}) = \mathbf{0}$  for  $j \notin \bar{\mathcal{E}}^+$ , we have  $\mathbf{x} = \sum_{j \in \bar{\mathcal{E}}^+} c_j \mathbf{a}_j$  for some  $\mathbf{a}_j \in \mathcal{M}_j$ . Then

$$\begin{aligned} \|\mathbf{x} - \Pi_{\bar{\mathcal{A}}}(\mathbf{x})\|_2^2 &= \left\| \sum_{j \in \bar{\mathcal{E}}^+} c_j \mathbf{a}_j - \sum_{j \in \bar{\mathcal{E}}^+} c_j \langle \mathbf{a}_j, \bar{\mathbf{a}}_j \rangle \bar{\mathbf{a}}_j \right\|_2^2 \\ &= \sum_{j \in \bar{\mathcal{E}}^+} c_j^2 \|\mathbf{a}_j - \langle \mathbf{a}_j, \bar{\mathbf{a}}_j \rangle \bar{\mathbf{a}}_j\|_2^2 \leq \sum_{j \in \bar{\mathcal{E}}^+} c_j^2 \|\mathbf{a}_j - \bar{\mathbf{a}}_j\|_2^2. \end{aligned}$$

Since  $\Pi_{\mathcal{M}_j}(\boldsymbol{\rho}) = \lambda \mathbf{a}_j, \Pi_{\mathcal{M}_j}(\bar{\boldsymbol{\rho}}) = \lambda \bar{\mathbf{a}}_j$ , we have

$$\|\mathbf{x} - \Pi_{\bar{\mathcal{A}}}(\mathbf{x})\|_2^2 \leq \frac{1}{\lambda^2} \sum_{j \in \bar{\mathcal{E}}^+} c_j^2 \|\Pi_{\mathcal{M}_j}(\boldsymbol{\rho}) - \Pi_{\mathcal{M}_j}(\bar{\boldsymbol{\rho}})\|_2^2 \leq \frac{R^2}{\lambda^2} \|\boldsymbol{\rho} - \bar{\boldsymbol{\rho}}\|_2^2$$

as claimed.  $\square$

## 4 Proof of Lemma 6

**Lemma 6** (Optimality Condition). *For any matrix  $H$  satisfying CNSC- $\mathcal{T}$ , the update*

$$\Delta \mathbf{x} = \underset{\mathbf{d}}{\text{argmin}} \quad h(\mathbf{x} + \mathbf{d}) + \mathbf{g}(\mathbf{x})^T \mathbf{d} + \frac{1}{2} \|\mathbf{d}\|_H^2 \quad (4)$$

has

$$F(\mathbf{x} + t\Delta \mathbf{x}) - F(\mathbf{x}) \leq -t \|\Delta \mathbf{z}\|_H^2 + O(t^2), \quad (5)$$

where  $\Delta \mathbf{z} = \Pi_{\mathcal{T}}(\Delta \mathbf{x})$ . Furthermore, if  $\mathbf{x}$  is an optimal solution,  $\Delta \mathbf{x} = \mathbf{0}$  satisfies (4).

*Proof.* By smoothness of  $f(\mathbf{x})$  and convexity of  $h(\mathbf{x})$ , we have

$$\begin{aligned} F(\mathbf{x} + t\Delta \mathbf{x}) - F(\mathbf{x}) &= h(\mathbf{x} + t\Delta \mathbf{x}) - h(\mathbf{x}) + f(\mathbf{x} + t\Delta \mathbf{x}) - f(\mathbf{x}) \\ &\leq t(h(\mathbf{x} + \Delta \mathbf{x}) - h(\mathbf{x})) + \mathbf{g}(\mathbf{x})^T (t\Delta \mathbf{x}) + \mathcal{O}(t^2). \end{aligned} \quad (6)$$

Then we try to bound the descent amount predicted by gradient  $t(h(\mathbf{x} + \Delta\mathbf{x}) - h(\mathbf{x}) + \mathbf{g}(\mathbf{x})^T \Delta\mathbf{x})$ . Since  $\Delta\mathbf{x}$  is optimal solution of (4), we have

$$\begin{aligned} & h(\mathbf{x} + \Delta\mathbf{x}) + \mathbf{g}(\mathbf{x})^T \Delta\mathbf{x} + \frac{1}{2} \|\Delta\mathbf{x}\|_H^2 \\ & \leq h(\mathbf{x} + t\Delta\mathbf{x}) + \mathbf{g}(\mathbf{x})^T (t\Delta\mathbf{x}) + \frac{1}{2} \|t\Delta\mathbf{x}\|_H^2 \\ & \leq th(\mathbf{x} + \Delta\mathbf{x}) + (1-t)h(\mathbf{x}) + \mathbf{g}(\mathbf{x})^T (t\Delta\mathbf{x}) + \frac{1}{2} \|t\Delta\mathbf{x}\|_H^2, \end{aligned} \quad (7)$$

which implies

$$(1-t)(h(\mathbf{x} + \Delta\mathbf{x}) - h(\mathbf{x})) + (1-t)\mathbf{g}(\mathbf{x})^T \Delta\mathbf{x} + \frac{1-t^2}{2} \|\Delta\mathbf{x}\|_H^2 \leq 0, \quad (8)$$

and therefore,

$$(h(\mathbf{x} + \Delta\mathbf{x}) - h(\mathbf{x})) + \mathbf{g}(\mathbf{x})^T \Delta\mathbf{x} \leq -\frac{1+t}{2} \|\Delta\mathbf{x}\|_H^2 = -\frac{1+t}{2} \|\Delta\mathbf{z}\|_H^2, \quad (9)$$

where  $\Delta\mathbf{z} = \Pi_{\mathcal{T}}(\Delta\mathbf{x})$  and last inequality follows from CNSC- $\mathcal{T}$  of  $H$ . Let  $t \rightarrow 1$  and combine (9) and (6), we obtain

$$F(\mathbf{x} + t\Delta\mathbf{x}) - F(\mathbf{x}) \leq -t\|\Delta\mathbf{z}\|_H^2 + \mathcal{O}(t^2), \quad (10)$$

which shows  $\Delta\mathbf{x}$  obtained from (4) is a descent direction if  $\Delta\mathbf{z} \neq \mathbf{0}$ .

Now suppose  $\mathbf{x}$  is an optimal solution of  $F(\mathbf{x})$ . Then the  $\Delta\mathbf{x}$  defined in (4) cannot be a descent direction, which means  $\Delta\mathbf{z}$  must be  $\mathbf{0}$ . However, since  $f(\mathbf{x})$  and  $H$  satisfy CNSC- $\mathcal{T}$ , when  $\Delta\mathbf{z} = \mathbf{0}$ , (4) reduced to

$$\Delta\mathbf{x} = \underset{\Delta\mathbf{y} \in \mathcal{T}^\perp}{\operatorname{argmin}} h(\mathbf{x} + \Delta\mathbf{y}). \quad (11)$$

$\Delta\mathbf{x} = \mathbf{0}$  satisfies (11) since  $\mathbf{x} = \mathbf{y} + \mathbf{z}$  is already a minimum of  $h(\mathbf{x}) + f(\mathbf{x})$ , while  $f(\mathbf{x})$  does not depend on  $\mathbf{y}$ , where  $\mathbf{y} = \Pi_{\mathcal{T}^\perp}(\mathbf{x})$ .  $\square$

## 5 Proof of Lemma 7

**Lemma 7.** *Suppose  $h(\mathbf{x})$  and  $f(\mathbf{x})$  are Lipchitz-continuous with Lipchitz constants  $L_h$  and  $L_f$ . In quadratic convergence phase (defined in Theorem 3), Proximal Newton Method has*

$$F(\mathbf{x}_t) - F(\bar{\mathbf{x}}) \leq L\|\mathbf{z}_t - \bar{\mathbf{z}}\|, \quad (12)$$

where  $L = \max\{L_h, L_f\}$  and  $\mathbf{z}_t = \Pi_{\mathcal{T}}(\mathbf{x}_t)$ ,  $\bar{\mathbf{z}} = \Pi_{\mathcal{T}}(\bar{\mathbf{x}})$ .

*Proof.* We prove (12) by showing that  $|f(\mathbf{z}_1) - f(\mathbf{z}_2)| \leq L_f\|\mathbf{z}_1 - \mathbf{z}_2\|$  and  $|h(\mathbf{z}_1 + \hat{\mathbf{y}}(\mathbf{z}_1)) - h(\mathbf{z}_2 + \hat{\mathbf{y}}(\mathbf{z}_2))| \leq L_h\|\mathbf{z}_1 - \mathbf{z}_2\|$  for any  $\mathbf{z}_1 \in \mathcal{T}$ ,  $\mathbf{z}_2 \in \mathcal{T}$ . Since  $f(\mathbf{z})$  does not depend on the null-component  $\mathbf{y}$ , the first inequality holds directly from the Lipchitz-continuity of  $f(\mathbf{z})$ . The second inequality holds since

$$h(\mathbf{z}_1 + \hat{\mathbf{y}}(\mathbf{z}_1)) \leq h(\mathbf{z}_1 + \hat{\mathbf{y}}(\mathbf{z}_2)) \leq h(\mathbf{z}_2 + \hat{\mathbf{y}}(\mathbf{z}_2)) + L_h\|\mathbf{z}_1 - \mathbf{z}_2\|$$

and

$$h(\mathbf{z}_2 + \hat{\mathbf{y}}(\mathbf{z}_2)) \leq h(\mathbf{z}_2 + \hat{\mathbf{y}}(\mathbf{z}_1)) \leq h(\mathbf{z}_1 + \hat{\mathbf{y}}(\mathbf{z}_1)) + L_h\|\mathbf{z}_1 - \mathbf{z}_2\|$$

by the definition of  $\hat{\mathbf{y}}(\mathbf{z}_1)$ ,  $\hat{\mathbf{y}}(\mathbf{z}_2)$  and Lipchitz-continuity of  $h(\mathbf{x})$ .  $\square$