8. Additional Demonstration Plots

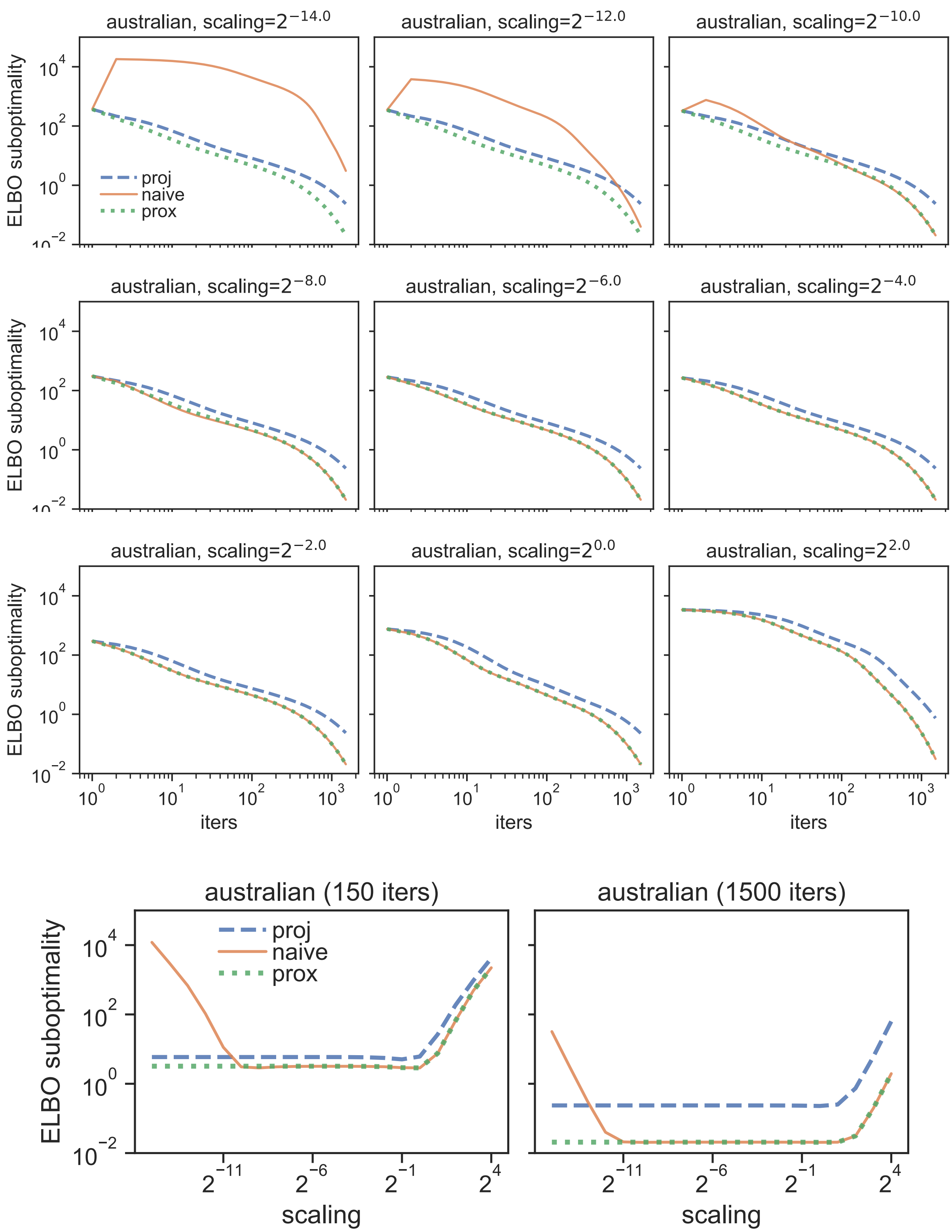

Figure 4. Looseness of the objective obtained by naive gradient descent ($\gamma = 1/M$), projected gradient descent ($\gamma = 1/(2M)$) and proximal gradient descent ($\gamma = 1/M$). Optimization starts with $\boldsymbol{m} = 0$ and $C = \rho I$ where ρ is a scaling factor.

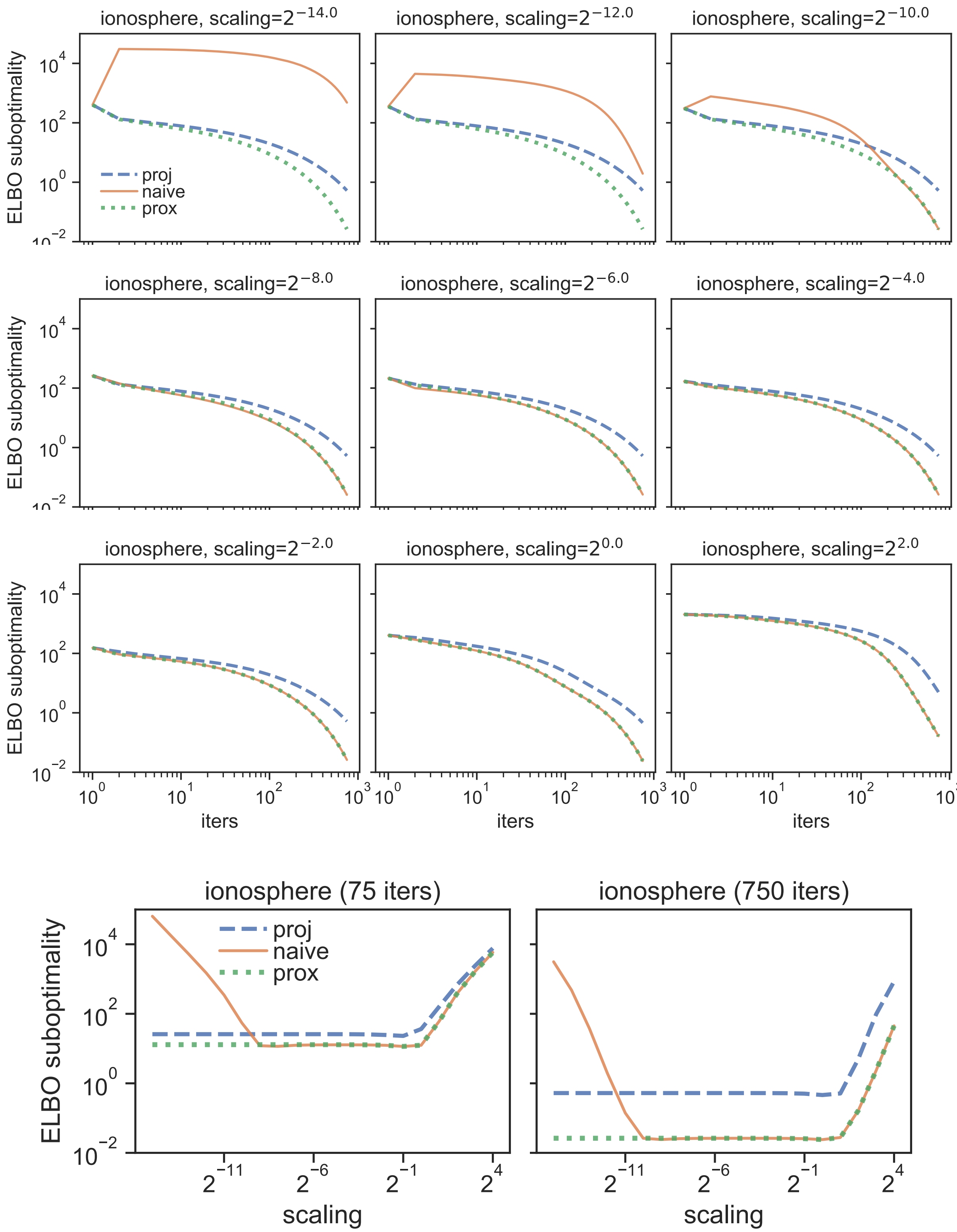

Figure 5. Looseness of the objective obtained by naive gradient descent ($\gamma = 1/M$), projected gradient descent ($\gamma = 1/(2M)$) and proximal gradient descent ($\gamma = 1/M$). Optimization starts with $\boldsymbol{m} = 0$ and $C = \rho I$ where ρ is a scaling factor.

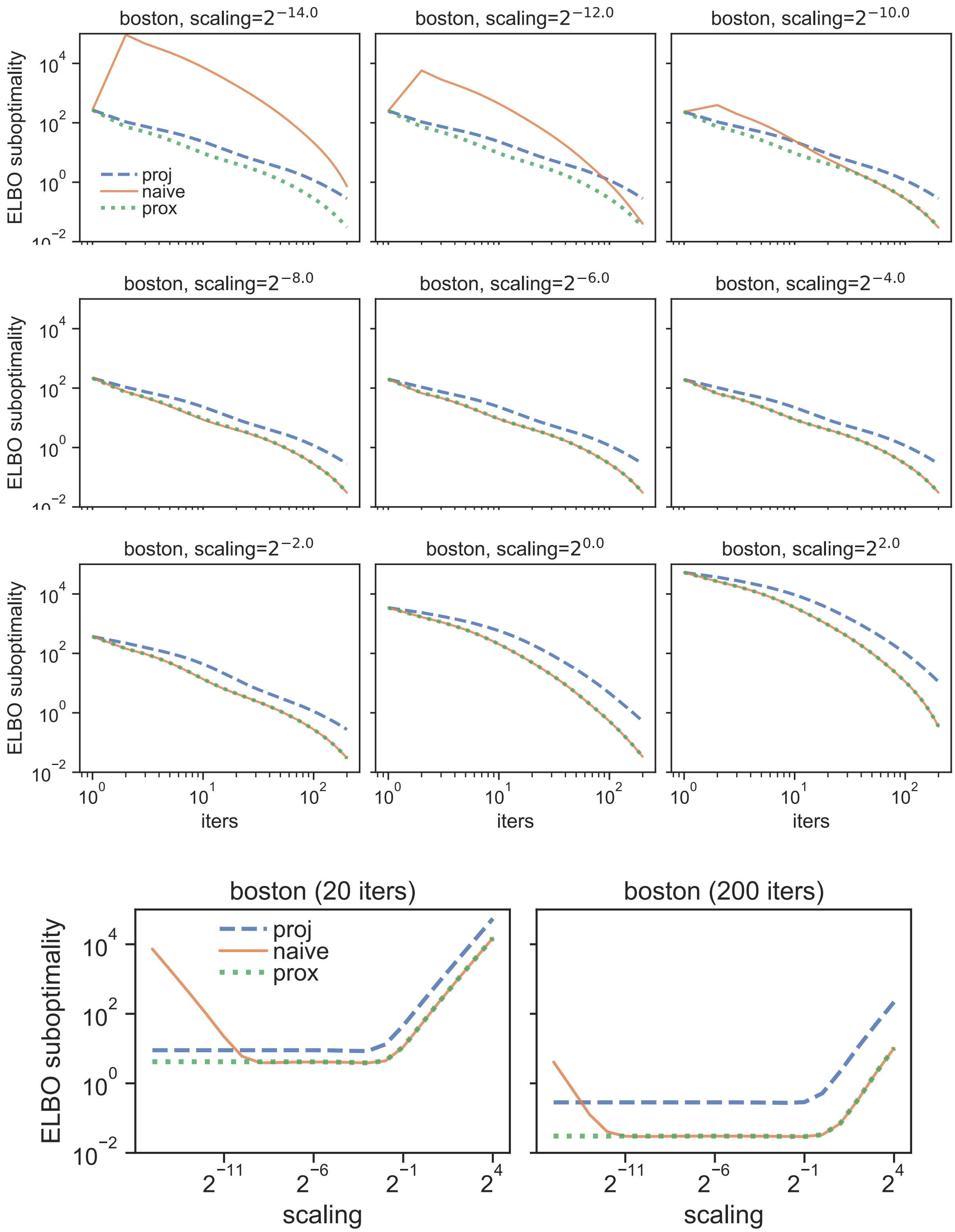

Figure 6. Looseness of the objective obtained by naive gradient descent ($\gamma = 1/M$), projected gradient descent ($\gamma = 1/(2M)$) and proximal gradient descent ($\gamma = 1/M$). Optimization starts with $\boldsymbol{m} = 0$ and $C = \rho I$ where ρ is a scaling factor.

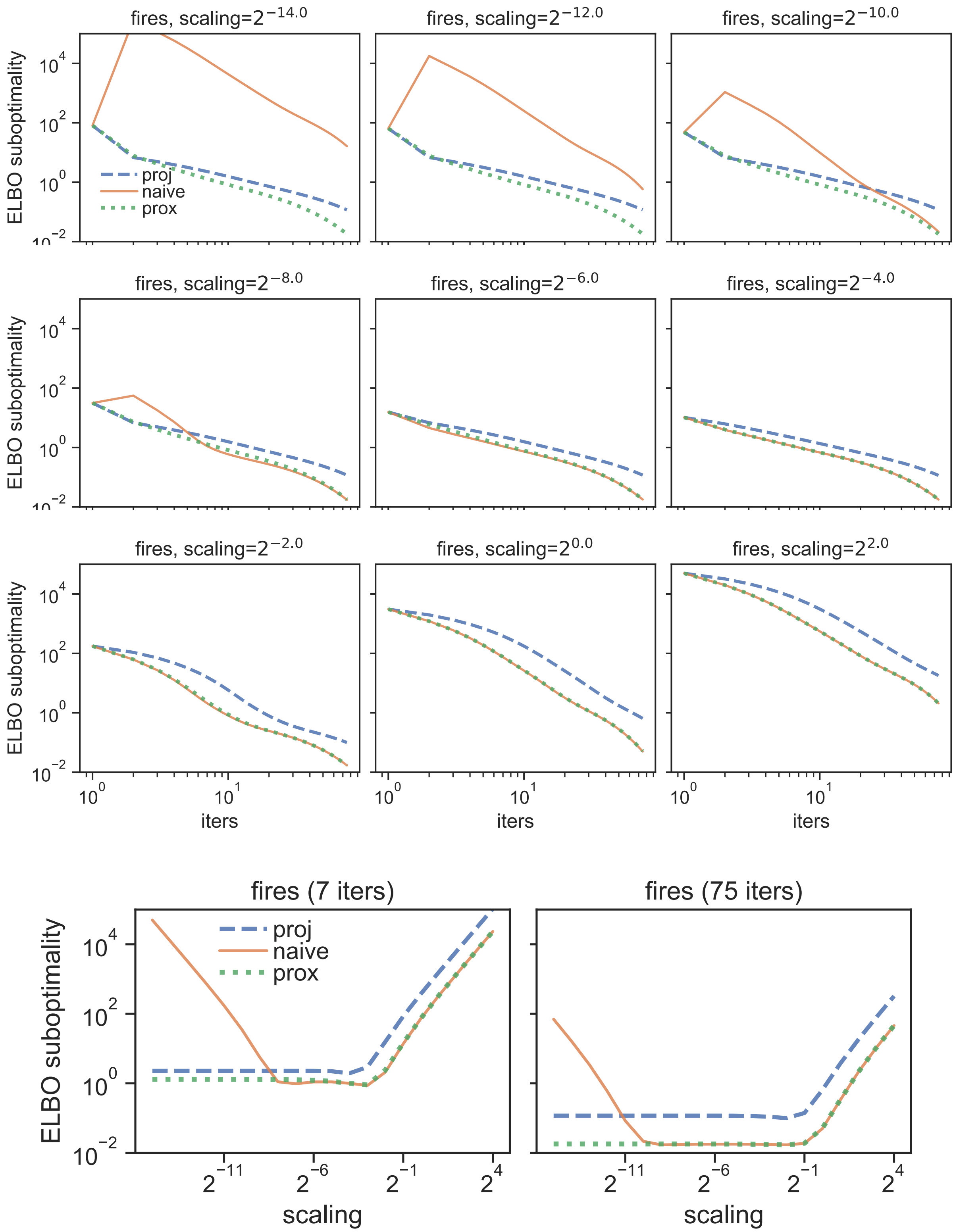

Figure 7. Looseness of the objective obtained by naive gradient descent ($\gamma = 1/M$), projected gradient descent ($\gamma = 1/(2M)$) and proximal gradient descent ($\gamma = 1/M$). Optimization starts with $\boldsymbol{m} = 0$ and $C = \rho I$ where ρ is a scaling factor.

9. Proofs for Technical Lemmas

This section gives proofs for the technical lemmas used in the main result. Firstly, we show that $\langle \cdot, \cdot \rangle_s$ is a valid inner-product.

Lemma 2. *$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_s = \mathbb{E}_{\mathsf{u} \sim s}\, \boldsymbol{a}(\mathsf{u})^\top \boldsymbol{b}(\mathsf{u})$ is a valid inner-product on squared-integrable $\boldsymbol{a} : \mathbb{R}^d \to \mathbb{R}^k$.*

Proof. The space of square integrable functions is $\left\{\boldsymbol{a} : \mathbb{R}^d \to \mathbb{R}^k \mid \mathbb{E}_{\mathsf{u} \sim s}\, a_i(\mathsf{u})^2 \leq \infty \; \forall i \in \{1, ..., k\}\right\}$. Since each component $a_i(\boldsymbol{u})$ and $b_i(\boldsymbol{u})$ is square-integrable with respect to $s(\boldsymbol{u})$ we know (by Cauchy-Schwarz) that $\mathbb{E}_{\mathsf{u} \sim s}\, a_i(\mathsf{u}) b_i(\mathsf{u}) \leq \sqrt{\mathbb{E}_{\mathsf{u} \sim s}\, a_i(\mathsf{u})^2}\sqrt{\mathbb{E}_{\mathsf{u} \sim s}\, b_i(\mathsf{u})}$ is finite and real. Therefore, we have by linearity of expectation that

$$\begin{aligned}\sum_{i=1}^k \mathop{\mathbb{E}}_{\mathsf{u} \sim s} a_i(\mathsf{u}) b_i(\mathsf{u}) &= \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \sum_{i=1}^k a_i(\mathsf{u}) b_i(\mathsf{u}) \\ &= \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \boldsymbol{a}(\mathsf{u})^\top \boldsymbol{b}(\mathsf{u}) \\ &= \langle \boldsymbol{a}, \boldsymbol{b} \rangle_s\end{aligned}$$

is finite and real for all $\boldsymbol{a}, \boldsymbol{b} \in V_s$. To show that $(V_s, \langle \cdot, \cdot \rangle_s)$ is a valid inner-product space, it is easy to establish all the necessary properties of the inner-product, namely for all $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in V_s$,

$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \boldsymbol{b}, \boldsymbol{a} \rangle$

$\langle \theta \boldsymbol{a}, \boldsymbol{b} \rangle = \theta \langle \boldsymbol{a}, \boldsymbol{b} \rangle$ for $\theta \in \mathbb{R}$

$\langle \boldsymbol{a} + \boldsymbol{b}, \boldsymbol{c} \rangle = \langle \boldsymbol{a}, \boldsymbol{c} \rangle + \langle \boldsymbol{b}, \boldsymbol{c} \rangle$

$\langle \boldsymbol{a}, \boldsymbol{a} \rangle \geq 0$

$\langle \boldsymbol{a}, \boldsymbol{a} \rangle = 0 \Leftrightarrow \boldsymbol{a} = \boldsymbol{0}$. (Where $\boldsymbol{0}(\boldsymbol{\varepsilon})$ is a function that always returns a vector of k zeros.) □

Next, we give three technical Lemmas, which do most of the work of the proof.

Lemma 3. *Let $\boldsymbol{a}_i(\boldsymbol{u}) = \frac{d}{dw_i} \boldsymbol{t_w}(\boldsymbol{u})$. This is independent of $\boldsymbol{w}$ and $\frac{dl(\boldsymbol{w})}{dw_i} = \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t_w} \rangle_s$.*

Proof. Now, we can write $l(\boldsymbol{w})$ as

$$l(\boldsymbol{w}) = \mathop{\mathbb{E}}_{\mathsf{z} \sim q_{\boldsymbol{w}}} f(\mathsf{z}) = \mathop{\mathbb{E}}_{\mathsf{u} \sim s} f\left(\boldsymbol{t_w}(\mathsf{u})\right).$$

Since $\boldsymbol{t_w}(\boldsymbol{u}) = C\boldsymbol{u} + \boldsymbol{m}$ is an affine function, it's easy to see that both $\frac{d}{dC_{ij}} \boldsymbol{t_w}(\boldsymbol{u})$ and $\frac{d}{d\boldsymbol{m}_i} \boldsymbol{t_w}(\boldsymbol{u})$ are independent of $\boldsymbol{w}$. Therefore, the gradient of $l(\boldsymbol{w})$ can be written as

$$\begin{aligned}\nabla_{w_i} l(\boldsymbol{w}) &= \nabla_{w_i} \mathop{\mathbb{E}}_{\mathsf{u} \sim s} f\left(\boldsymbol{t_w}(\mathsf{u})\right) \\ &= \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \nabla_{w_i} \boldsymbol{t_w}(\mathsf{u})^\top \nabla f\left(\boldsymbol{t_w}(\mathsf{u})\right). \\ &= \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t_w} \rangle_s.\end{aligned}$$

□

Lemma 4. *If s is standardized, then the functions $\{\boldsymbol{a}_i\}$ are orthonormal in $\langle \cdot, \cdot \rangle_s$.*

Proof. It is easy to calculate that

$$\begin{aligned}\frac{d}{dm_i} \boldsymbol{t_w}(\boldsymbol{u}) &= \boldsymbol{e}_i \\ \frac{d}{dC_{ij}} \boldsymbol{t_w}(\boldsymbol{u}) &= \boldsymbol{e}_i u_j,\end{aligned}$$

where $\boldsymbol{e}_i$ is the indicator vector in the i-th component. Therefore, we have that

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathsf{u} \sim s} \left(\frac{d}{dm_i} \boldsymbol{t_w}(\mathsf{u})\right)^\top \left(\frac{d}{dm_j} \boldsymbol{t_w}(\mathsf{u})\right) \\ &= \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \boldsymbol{e}_i^\top \boldsymbol{e}_j \\ &= I[i = j]\end{aligned}$$

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathsf{u} \sim s} \left(\frac{d}{dC_{ij}} \boldsymbol{t_w}(\mathsf{u})\right)^\top \left(\frac{d}{dm_k} \boldsymbol{t_w}(\mathsf{u})\right) \\ &= \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \mathsf{u}_j \boldsymbol{e}_i^\top \boldsymbol{e}_k \\ &= I[i = k] \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \mathsf{u}_j \\ &= 0\end{aligned}$$

(since zero mean)

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathsf{u} \sim s} \left(\frac{d}{dC_{ij}} \boldsymbol{t_w}(\mathsf{u})\right)^\top \left(\frac{d}{dC_{kl}} \boldsymbol{t_w}(\mathsf{u})\right) \\ &= \mathbb{E}\, \mathsf{u}_j \mathsf{u}_l \boldsymbol{e}_i^\top \boldsymbol{e}_k \\ &= I[i = k] \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \mathsf{u}_j \mathsf{u}_l \\ &= I[i = k] I[j = l]\end{aligned}$$

(since unit variance and zero mean)

These three identities are equivalent to stating that $\{\boldsymbol{a}_i\}$ are orthonormal in $\langle \cdot, \cdot \rangle_s$. □

Lemma 5. *If s is standardized, then $\mathbb{E}_{\mathsf{u} \sim s} \|\boldsymbol{t_w}(\mathsf{u}) - \boldsymbol{t_v}(\mathsf{u})\|_2^2 = \|\boldsymbol{w} - \boldsymbol{v}\|_2^2$.*

Proof. Let $\Delta \boldsymbol{m}$ and ΔS denote the difference of the $\boldsymbol{m}$ and S parts of $\boldsymbol{w}$, respectively. We want to calculate

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathsf{u} \sim s} \|\boldsymbol{t_w}(\mathsf{u}) - \boldsymbol{t_v}(\mathsf{u})\|_2^2 \\ &= \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \|\Delta C \boldsymbol{\varepsilon} + \Delta \boldsymbol{m}\|_2^2 \\ &= \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \left(\|(\Delta C)\mathsf{u}\|_2^2 + 2\Delta \boldsymbol{m}^\top \Delta C \mathsf{u} + \|\Delta \boldsymbol{m}\|_2^2\right).\end{aligned}$$

It is easy to see that the expectation of the middle term is zero, and the last is a constant. The expectation of the first term is

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathsf{u}\sim s} \|(\Delta C)\mathsf{u}\|_2^2 &= \mathop{\mathbb{E}}_{\mathsf{u}\sim s} \mathsf{u}^\top (\Delta C)^\top (\Delta C)\mathsf{u} \\
&= \mathop{\mathbb{E}}_{\mathsf{u}\sim s} \operatorname{tr}\left(\mathsf{u}^\top (\Delta C)^\top (\Delta C)\mathsf{u}\right) \\
&= \mathop{\mathbb{E}}_{\mathsf{u}\sim s} \operatorname{tr}\left((\Delta C)^\top (\Delta C)\mathsf{u}\mathsf{u}^\top\right) \\
&= \operatorname{tr}\left((\Delta C)^\top (\Delta C)\right) = \|\nabla C\|_F^2 . \\
&\quad \text{(since zero mean and unit variance)}
\end{aligned}$$

Putting this together gives that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathsf{u}\sim s} \|\boldsymbol{t}_{\boldsymbol{w}}(\mathsf{u}) - \boldsymbol{t}_{\boldsymbol{v}}(\mathsf{u})\|_2^2 &= \|\Delta C\|_F^2 + \|\Delta \boldsymbol{m}\|_2^2 \\
&= \|\boldsymbol{w} - \boldsymbol{v}\|_2^2 .
\end{aligned}$$

□

10. Proof for Example Function

Theorem 6. *Let* $q_{\boldsymbol{w}} = \mathrm{LocScale}(\boldsymbol{m}, C, s)$ *with parameters* $\boldsymbol{w} = (\boldsymbol{m}, C)$ *and a standardized base distribution* s *and let* $f(\boldsymbol{z}) = \frac{a}{2}\left\|\boldsymbol{z} - \boldsymbol{z}^*\right\|_2^2$. *Then* $l(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z}\sim q_{\boldsymbol{w}}} f(\mathsf{z}) = \frac{a}{2}(\|\boldsymbol{m} - \boldsymbol{z}^*\|_2^2 + \|C\|_F^2)$.

Proof. For a general distribution, we have that

$$
\begin{aligned}
\mathbb{E} f(\mathbf{z}) &= \frac{a}{2}\,\mathbb{E}\left\|\mathsf{z} - \mathbb{E}[\mathsf{z}] + \mathbb{E}[\mathsf{z}] - \boldsymbol{z}^*\right\|_2^2 \\
&= \frac{a}{2}\,\mathbb{E}\Big(\left\|\mathsf{z} - \mathbb{E}[\mathsf{z}]\right\|_2^2 \\
&\quad + 2\left(\mathsf{z} - \mathbb{E}[\mathsf{z}]\right)^\top\left(\mathbb{E}[\mathsf{z}] - \boldsymbol{z}^*\right) + \left\|\mathbb{E}[\mathsf{z}] - \boldsymbol{z}^*\right\|_2^2\Big) \\
&= \frac{a}{2}\left(\mathrm{tr}\,\mathbb{V}[\mathsf{z}] + \left\|\mathbb{E}[\mathsf{z}] - \boldsymbol{z}^*\right\|_2^2\right).
\end{aligned}
$$

Now, if $q_{\boldsymbol{w}}$ is a location-scale family, we have that $\mathsf{z} = C\mathsf{u} + \boldsymbol{m}$. Thus,

$$
\begin{aligned}
\mathrm{tr}\,\mathbb{V}[\mathsf{z}] &= \mathrm{tr}\,\mathbb{V}[C\mathsf{u} + \boldsymbol{m}] \\
&= \mathrm{tr}\,\mathbb{V}[C\mathsf{u}] \\
&= \mathrm{tr}\,C\,\mathbb{V}[\mathsf{u}]C^\top \\
&= \mathrm{tr}\,CC^\top\,\mathbb{V}[\mathsf{u}].
\end{aligned}
$$

Meanwhile, we have that

$$
\begin{aligned}
\left\|\mathbb{E}[\mathsf{z}] - \boldsymbol{z}^*\right\|_2^2 &= \left\|\mathbb{E}[C\mathsf{u} + \boldsymbol{m}] - \boldsymbol{z}^*\right\|_2^2 \\
&= \left\|C\,\mathbb{E}[\mathsf{u}] + \boldsymbol{m} - \boldsymbol{z}^*\right\|_2^2
\end{aligned}
$$

Thus,

$$
\mathbb{E} f(\mathbf{z}) = \frac{a}{2}\left(\mathrm{tr}\,C\,\mathbb{V}[\mathsf{u}]C^\top + \left\|C\,\mathbb{E}[\mathsf{u}] + \boldsymbol{m} - \boldsymbol{z}^*\right\|_2^2\right).
$$

The case where s is standardized follows from substituting $\mathbb{E}[\mathsf{u}] = 0$ and $\mathbb{V}[\mathsf{u}] = I$ and applying the fact that $\mathrm{tr}\,CC^\top = \|C\|_F^2$. $\square$

11. Proofs for Solution Guarantees

Lemma 8. *Let* $q_{\boldsymbol{w}} = \mathrm{LocScale}(\boldsymbol{m}, C, s)$ *with parameters* $\boldsymbol{w} = (\boldsymbol{m}, C)$ *and a standardized and spherically symmetric base distribution* s. *Let* $l(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z}\sim q_{\boldsymbol{w}}} f(\mathsf{z})$. *Suppose* C *is diagonal and* f *is* M*-smooth. Then,* $|\frac{dl(\boldsymbol{w})}{dC_{ii}}| \leq M|C_{ii}|$.

Proof. Define $\boldsymbol{w}'$ to be $\boldsymbol{w}$ but with C_{ii} set to zero. We will first show that $\frac{dl(\boldsymbol{w}')}{dC_{ii}} = 0$. Using the definition of $\boldsymbol{t}_w$ and the fact that $\frac{d}{dC_{ij}} \boldsymbol{t}_w(u) = \boldsymbol{e}_i u_j$ gives that

$$\begin{aligned}
\frac{d}{dC_{ii}} l(\boldsymbol{w}') &= \mathop{\mathbb{E}}_{\mathsf{u}\sim s} \frac{d}{dC_{ii}} f(\boldsymbol{t}_{\boldsymbol{w}'}(\mathsf{u})) && (11)\\
&= \mathop{\mathbb{E}}_{\mathsf{u}\sim s} \mathsf{u}_i \boldsymbol{e}_i^\top \nabla f(\boldsymbol{t}_{\boldsymbol{w}'}(\mathsf{u})) && (12)\\
&= 0. && (13)
\end{aligned}$$

The final equality above follows from the facts that $\mathbb{E}\, \mathsf{u}_i = 0$ and $\mathsf{u}_i \perp \boldsymbol{e}_i^\top \nabla f(\boldsymbol{t}_{\boldsymbol{w}'}(\mathsf{u}))$ (Since $\boldsymbol{t}_{\boldsymbol{w}'}(\boldsymbol{u})$ ignores $\boldsymbol{u}_i$) so the expectation in Eq. (11) is over two independent random variables, one with mean zero. Now, by Thm. 1, l is also M-smooth, thus

$$\begin{aligned}
\left|\frac{dl(\boldsymbol{w})}{dC_{ii}}\right| &= \left|\frac{dl(\boldsymbol{w}')}{dC_{ii}} - \frac{dl(\boldsymbol{w})}{dC_{ii}}\right|\\
&\leq \|\nabla l(\boldsymbol{w}') - \nabla l(\boldsymbol{w})\|_2\\
&\leq M \|\boldsymbol{w}' - \boldsymbol{w}\|_2\\
&= M |C_{ii}|.
\end{aligned}$$

□

Theorem 7. *Let* $q_{\boldsymbol{w}} = \mathrm{LocScale}(\boldsymbol{m}, C, s)$ *with parameters* $\boldsymbol{w} = (\boldsymbol{m}, C)$ *and a standardized and spherically symmetric base distribution* s. *Suppose* $\boldsymbol{w}$ *minimizes* $l(\boldsymbol{w}) + h(\boldsymbol{w})$ *from Eq. (1) and* $\log p(\boldsymbol{z}, \boldsymbol{x})$ *is* M*-smooth over* $\boldsymbol{z}$. *Then,* $\boldsymbol{w} \in \mathcal{W}_M$.

Proof. First, suppose that C is diagonal. Since $\boldsymbol{w}$ minimizes $l + h$, $\nabla l(\boldsymbol{w}) = -\nabla h(\boldsymbol{w})$. The gradient of h with respect to C is $-C^{-\top}$. Thus, $|\frac{dl(\boldsymbol{w})}{dC_{ii}}| = |\frac{dh(\boldsymbol{w})}{dC_{ii}}| = \frac{1}{|C_{ii}|}$. But by Lem. 8, $|\frac{dl(\boldsymbol{w})}{dC_{ii}}| \leq M|C_{ii}|$. This establishes the claim for diagonal C.

Now, consider some non-diagonal C. Let the singular value decomposition be $C = USV^\top$. Define $f_U(\boldsymbol{z}) = f(U\boldsymbol{z})$ and define l_U with respect to f_U. Let $\boldsymbol{w}' = (S, U^\top \boldsymbol{m})$. Then, the following statements are equivalent to $\boldsymbol{w} \in \mathrm{argmin}_{\boldsymbol{w}}\, l(\boldsymbol{w}) + h(\boldsymbol{w})$:

$$\begin{aligned}
&(C, \boldsymbol{m}) \in \mathop{\mathrm{argmin}}_{(C,\boldsymbol{m})} \mathop{\mathbb{E}}_{\mathsf{u}\sim s} f(C\mathsf{u} + \boldsymbol{m}) - \log|C|\\
\Leftrightarrow\ &(S, \boldsymbol{m}) \in \mathop{\mathrm{argmin}}_{(S,\boldsymbol{m})} \mathop{\mathbb{E}}_{\mathsf{u}\sim s} f\left(USV^\top \mathsf{u} + \boldsymbol{m}\right) - \log\left|USV^\top\right|\\
\Leftrightarrow\ &(S, \boldsymbol{m}) \in \mathop{\mathrm{argmin}}_{(S,\boldsymbol{m})} \mathop{\mathbb{E}}_{\mathsf{u}\sim s} f(US\mathsf{u} + \boldsymbol{m}) - \log|S|\\
\Leftrightarrow\ &(S, \boldsymbol{m}) \in \mathop{\mathrm{argmin}}_{(S,\boldsymbol{m})} \mathop{\mathbb{E}}_{\mathsf{u}\sim s} f_U\left(S\mathsf{u} + U^\top \boldsymbol{m}\right) - \log|S|\\
\Leftrightarrow\ &\boldsymbol{w}' \in \mathop{\mathrm{argmin}}_{\boldsymbol{w}} l_U(\boldsymbol{w}) + h(\boldsymbol{w}).
\end{aligned}$$

Thus, $\boldsymbol{w}$ minimizing $l + h$ is equivalent to $\boldsymbol{w}'$ minimizing $l_U + h$. Since f_U is M-smooth and S is diagonal, we know that $S_{ii} \geq \frac{1}{\sqrt{M}}$ for all . □

12. Proofs with Convexity

Theorem 10. *Let* $q_{\boldsymbol{w}} = \mathrm{LocScale}(\boldsymbol{m}, C, s)$ *with parameters* $\boldsymbol{w} = (\boldsymbol{m}, C)$ *and a standardized and spherically symmetric base distribution* s*. Suppose* $\boldsymbol{w}$ *minimizes* $l(\boldsymbol{w}) + h(\boldsymbol{w})$ *from Eq. (1) and* $-\log p(\boldsymbol{z}, \boldsymbol{x})$ *is* c*-strongly convex over* $\boldsymbol{z}$*. Then,* $\|C\|_F^2 + \|\boldsymbol{m} - \boldsymbol{z}^*\|_2^2 \leq \frac{d}{c}$*, where* $\boldsymbol{z}^* = \mathrm{argmax}_{\boldsymbol{z}} \log(\boldsymbol{z}, \boldsymbol{x})$.

It's easy to see that l is minimized by $\bar{\boldsymbol{w}} = (\boldsymbol{z}^*, \mathbf{0}_{d\times d})$. By Thm. 9, $l(\boldsymbol{w})$ is c-strongly convex. Thus applying a standard inner-product result on strong convexity (Nesterov, 2014, Thm. 2.1.9),

$$
\begin{aligned}
c \|\boldsymbol{w} - \bar{\boldsymbol{w}}\|_2^2 &\leq \langle \nabla l(\boldsymbol{w}) - \nabla l(\bar{\boldsymbol{w}}), \boldsymbol{w} - \bar{\boldsymbol{w}} \rangle \\
&\quad \text{(since } l \text{ is strongly convex)} \\
&= \langle \nabla l(\boldsymbol{w}), \boldsymbol{w} - \bar{\boldsymbol{w}} \rangle \\
&\quad \text{(since } \nabla l(\bar{\boldsymbol{w}}) = 0) \\
&= - \langle \nabla h(\boldsymbol{w}), \boldsymbol{w} - \bar{\boldsymbol{w}} \rangle \\
&\quad \text{(since } \nabla l(\boldsymbol{w}) + \nabla h(\boldsymbol{w}) = 0) \\
&= \mathrm{tr}\left(C^{-\top} C\right) \\
&\quad \text{(since } \nabla_C h(\boldsymbol{w}) = -C^{-\top}, \nabla_{\boldsymbol{m}} h(\boldsymbol{w}) = 0). \\
&= \mathrm{tr}\, I = d.
\end{aligned}
$$

The result follows from observing that $\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|_2^2 = \|C\|_F^2 + \|\boldsymbol{m} - \boldsymbol{z}^*\|_2^2$.

13. Convergence Considerations

Lemma 12. *Let* $q_{\boldsymbol{w}} = \mathrm{LocScale}(\boldsymbol{m}, C, s)$ *with parameters* $\boldsymbol{w} = (\boldsymbol{m}, C)$*. Then,* $h(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z}\sim q_{\boldsymbol{w}}} [\log q_{\boldsymbol{w}}(\mathsf{z})]$ *is* M*-smooth over* $\mathcal{W}_M$*.*

Proof. Take $\boldsymbol{w} = (C, \boldsymbol{m}) \in \mathcal{W}_M$ and $\boldsymbol{v} = (B, \boldsymbol{n}) \in \mathcal{W}_M$. We write $h(C)$ since $h(\boldsymbol{w})$ is independent of $\boldsymbol{m}$. The gradient is $\nabla h(C) = C^{-T}$. Now, use that $\|AX\|_F \leq \|A\|_2 \|X\|_F$ to get that

$$\begin{aligned}\|\nabla h(B) - \nabla h(C)\|_F &= \left\|B^{-1} - C^{-1}\right\|_F \\ &= \left\|B^{-1}(B - C)C^{-1}\right\|_F \\ &\leq \left\|B^{-1}\right\|_2 \left\|C^{-1}\right\|_2 \|B - C\|_F .\end{aligned}$$

But, since $\boldsymbol{w} \in \mathcal{W}_M$, $\left\|C^{-1}\right\|_2 = \frac{1}{\sigma_{\min}(C)} \leq \sqrt{M}$ and similarly for C. This establishes that $\|\nabla h(B) - \nabla h(C)\|_F \leq M \|B - C\|_F$, equivalent to the result. ☐

Theorem 13. *Suppose* $h(\boldsymbol{w})$ *corresponds to a location-scale family with a standardized* s*, and* $\boldsymbol{w} = (\boldsymbol{m}, C)$*.*

- *If* C *has singular value decomposition* $C = USV^\top$*, then* $\mathrm{proj}_{\mathcal{W}_M}(\boldsymbol{w}) = (\boldsymbol{m}, UTV^\top)$*, where* T *is a diagonal matrix with* $T_{ii} = \max\left(S_{ii}, \frac{1}{\sqrt{M}}\right)$*.*
- *If* C *is triangular with a positive diagonal, then* $\mathrm{prox}_\gamma(\boldsymbol{w}) = (\boldsymbol{m}, C + \Delta C)$*, where* ΔC *is a diagonal matrix with* $\Delta C_{ii} = \frac{1}{2}\left(\sqrt{C_{ii}^2 + 4\gamma} - C_{ii}\right)$*.*

Proof. **(Proximal Operator)** We know that $h(\boldsymbol{w}) = \mathrm{Const.} - \log|C|$. Write $\boldsymbol{w} = (\boldsymbol{m}, C)$ and $\boldsymbol{v} = (\boldsymbol{n}, B)$. Then, we can write the proximal operator as

$$\underset{\lambda}{\mathrm{prox}}(\boldsymbol{w}) = \underset{\boldsymbol{v}}{\mathrm{argmin}} - \log|B| + \frac{1}{2\lambda}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2$$

Now, assuming that C is triangular, the solution will leave all entries of $\boldsymbol{w}$ other than the diagonal entries of C unchanged. Then, we will have that $\log|B| = \sum_{i=1}^d \log B_{ii}$. Since

$$\underset{x>0}{\mathrm{argmin}} - \log x + \frac{1}{2\lambda}(x - y)^2 \quad = \quad \frac{y + \sqrt{y^2 + 4\lambda}}{2}$$

The solution is to set

$$\begin{aligned}B_{ii} &= \frac{1}{2}\left(C_{ii} + \sqrt{C_{ii}^2 + 4\lambda}\right) \\ &= C_{ii} + \frac{1}{2}\left(\sqrt{C_{ii}^2 + 4\lambda} - C_{ii}\right).\end{aligned}$$

(Projection Operator) Von-Neumann's trace inequality states that $\left|\mathrm{tr}\, A^\top B\right| \leq \sum_i \sigma_i(A)\sigma_i(B)$. Consider any candidate solution B with SVD $QTP^\top$. Then, we can write that

$$\begin{aligned}\|B - C\|_F^2 &= \mathrm{tr}\,(B - C)^\top (B - C) \\ &= \|B\|_F^2 - 2\,\mathrm{tr}(B^\top C) + \|C\|_F^2 \\ &\geq \|T\|_F^2 - 2\sum_i T_{ii}S_{ii} + \sum_i S_{ii}^2 \\ &= \sum_i (T_{ii} - S_{ii})^2 .\end{aligned}$$

We can minimize this lower bound by choosing $T_{ii} = \max(1/\sqrt{M}, S_{ii})$, with a corresponding value of $\sum_i \max(0, 1/\sqrt{M} - S_{ii})^2$. Thus any valid solution will have $\|B - C\|_F^2$ at least this large.

However, suppose we choose $B = UT_{ii}V^\top$ with T_{ii} as above. Then,

$$\|B - C\|_F^2 = \left\|UTV^\top - USV^\top\right\|_F^2 = \sum_i (T_{ii} - S_{ii})^2,$$

so this value B is optimal. ☐