
Nonconvex Stochastic Scaled Gradient Descent and Generalized Eigenvector Problems (Supplementary Material)

Chris Junchi Li¹

Michael I. Jordan^{1,2}

¹Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, California, USA

²Department of Statistics, UC Berkeley, Berkeley, California, USA

A PROBLEM-DEPENDENT PARAMETERS FOR GEV

We need to verify that the objective function for the GEV problems is indeed in the class of strict-saddle functions. For the GEV problem, the objective function of interest is

$$F(\mathbf{v}) = -\frac{\mathbf{v}^\top \mathbf{A} \mathbf{v}}{\mathbf{v}^\top \mathbf{B} \mathbf{v}}, \quad \text{such that } c(\mathbf{v}) = \|\mathbf{v}\|^2 - 1, \quad (1)$$

where \mathbf{A} and \mathbf{B} are two symmetric matrices. We make one additional mild assumption on the eigenstructure of matrices \mathbf{A} and \mathbf{B} .

Assumption 1 *The matrix $\mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{-1/2}$ is diagonalizable with eigenvalues $\lambda_1 > \lambda_2 > \dots > \lambda_d$. Moreover, $\lambda_{\min}(\mathbf{B}) > 0$.*

As our argument proceeds, one can safely assume $\mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{-1/2}$ being diagonal without loss of generality, so we will proceed with such. Under Assumption 1 we denote the minimal gap of λ_i 's as

$$\lambda_{\text{gap}} = \min_{1 \leq i \leq d-1} (\lambda_i - \lambda_{i+1}) > 0. \quad (2)$$

We prove that under the mild Assumption 1, the objective function for generalized eigenvector problem is strict-saddle as in Definition ?? if the parameters are chosen properly:

For the generalized eigenvector problem, the objective function of interest is

$$F(\mathbf{v}) = -\frac{\mathbf{v}^\top \mathbf{A} \mathbf{v}}{\mathbf{v}^\top \mathbf{B} \mathbf{v}}, \quad \text{such that } c(\mathbf{v}) = \|\mathbf{v}\|^2 - 1 = 0, \quad (3)$$

where \mathbf{A} and \mathbf{B} are two real symmetric matrices with \mathbf{B} being strictly positive-definite. In the following lemma, we verify that the objective function $F(\mathbf{v})$ in (3) satisfies Assumption ??; that is, $D(\mathbf{v})$, $F(\mathbf{v})$, $\nabla F(\mathbf{v})$, $\nabla^2 F(\mathbf{v})$ are Lipschitz continuous within $\{\mathbf{v} : \|\mathbf{v}\| \leq 1, \|\mathbf{v} - \mathbf{v}^*\| \leq \delta\}$.

Proposition 1 *Assumption ?? holds for $F(\mathbf{v})$ in GEV problem (3) with constants*

$$L_D = 2\|\mathbf{B}\|^2, \quad L_F = \frac{4\|\mathbf{A}\|\|\mathbf{B}\|}{(1-\delta)^2\lambda_{\min}^2(\mathbf{B})}, \quad L_K = \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}, \quad L_Q = \frac{232\|\mathbf{A}\|\|\mathbf{B}\|^3}{(1-\delta)^4\lambda_{\min}^4(\mathbf{B})}.$$

The proof of Proposition 1 is deferred to §B.3. With the Lipschitz parameters given above, we consider the initialization condition (??). The neighborhood radius on the right-hand side of (??) can be viewed as a function of δ that is maximized at some $\delta^* \in (0, 1)$, when all other constants are fixed. The region covered in the local convergence analysis is maximized with such a choice of δ^* .

Proposition 2 Under Assumption 1, the only local minimizers of (3) are $\pm \mathbf{e}_1$, and the function satisfies the $(\mu, \beta, \gamma, \delta)$ -strict saddle condition for

$$\begin{aligned}\mu &= (\lambda_1 - \lambda_2) \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|}, & \beta &= (\lambda_1 - \lambda_2) \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|}, \\ \gamma &= \lambda_{\text{gap}}^3 \frac{\lambda_{\min}^8(\mathbf{B})}{(8)84^2 \|\mathbf{A}\|^2 \|\mathbf{B}\|^6}, & \delta &= (\lambda_1 - \lambda_2) \frac{\lambda_{\min}^4(\mathbf{B})}{168 \|\mathbf{A}\| \|\mathbf{B}\|^3}.\end{aligned}\tag{4}$$

To verify the strict-saddle parameters and conclude Proposition 2, we first conclude the parameters for the objective function of the eigenvector problem:

Lemma 3 Under Assumption 1, and with the choices of parameters as in (4), we have the following:

(i) Suppose $\|g(\mathbf{x})\| \leq \gamma$ and $|\mathbf{e}_1^\top \mathbf{B}^{1/2} \mathbf{x}| \leq (1/2) \|\mathbf{B}^{1/2} \mathbf{x}\|$. Let the vector

$$\mathbf{v} \equiv \frac{P_{\mathcal{T}(\mathbf{x})} \mathbf{B}^{-1/2} \mathbf{e}_1}{\|P_{\mathcal{T}(\mathbf{x})} \mathbf{B}^{-1/2} \mathbf{e}_1\|},$$

then $\mathbf{v} \in \mathcal{T}(\mathbf{x})$, $\|\mathbf{v}\| = 1$, and we have

$$\mathbf{v}^\top \mathcal{H}(\mathbf{x}) \mathbf{v} \leq -\beta.\tag{5}$$

(ii) Suppose $\|g(\mathbf{x})\| \leq \gamma$ and $|\mathbf{e}_1^\top \mathbf{B}^{1/2} \mathbf{x}| > (1/2) \|\mathbf{B}^{1/2} \mathbf{x}\|$. Then there is a local minimizer \mathbf{x}^* such that $\|\mathbf{x} - \mathbf{x}^*\| \leq \delta$, and for all $\mathbf{x}' \in \mathbf{B}_{2\delta}(\mathbf{x}^*)$ we have for all $\hat{\mathbf{v}} \in \mathcal{T}(\mathbf{x}')$ and $\|\hat{\mathbf{v}}\| = 1$

$$\hat{\mathbf{v}}^\top \mathcal{H}(\mathbf{x}') \hat{\mathbf{v}} \geq \mu.\tag{6}$$

It is straightforward from Definition ?? of strict-saddle property that Lemma 3 leads to Proposition 2 immediately. We postpone the details to §D. Intuitively, the parameters are only dependent on the differences of the consecutive eigenvalues $\lambda_1 - \lambda_2, \dots, \lambda_{d-1} - \lambda_d$, since we can always add each eigenvalue λ_i by an arbitrary constant and keep the constrained optimization problem (3) unchanged. We also remark that restricted to our analysis, the parameters in (4) might not be the sharpest possible choices. However, we do provide, to the best of our knowledge, a first identification of strict-saddle parameters for the GEV problem, and hence Theorems ?? and ?? apply.

B PROOFS

In this section, we provide detailed proofs of our main results.

B.1 PROOF OF PROPOSITION ??

This subsection provides a proof for Proposition ?? on the convergence to a local minimizer. Under the initialization condition (??), there exists a local minimizer $\mathbf{v}^* \in \mathbf{B}_\delta(\mathbf{v}_0)$ of $F(\mathbf{v})$ such that $\mathbf{u}^\top \mathcal{H}(\mathbf{v}^*) \mathbf{u} \geq \mu \|\mathbf{u}\|^2$ for all $\mathbf{u} \in \mathcal{T}(\mathbf{v}^*)$.

For a positive quantity M to be determined later, let

$$\mathcal{T}_M = \inf \{t \geq 1 : \|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\| > M\}.\tag{7}$$

In words, \mathcal{T}_M is the first t such that the norm of the stochastic scaled-gradient $\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)$ exceeds M . We first provide the following lemma.

Lemma 4 Assume all conditions in Theorem ?. For any positive ϵ , let

$$M = \mathcal{V} \log^{1/\alpha} \epsilon^{-1}.\tag{8}$$

Then, we have

$$\mathbb{P}(\mathcal{T}_M \leq T_\eta^*) \leq 2T_\eta^* \epsilon.$$

The proof of Lemma 4 is a straightforward corollary of a union bound and Assumption ??, and is provided in §E.1.

Recall the definitions of the manifold gradient $g(\mathbf{v})$ and the Hessian $\mathcal{H}(\mathbf{v})$ in (??) and (??). Under a unit spherical constraint $c(\mathbf{v}) = \|\mathbf{v}\|^2 - 1 = 0$, their definitions simplify to

$$g(\mathbf{v}) = (\mathbf{I} - \mathbf{v}\mathbf{v}^\top) \nabla F(\mathbf{v}) \quad \text{and} \quad \mathcal{H}(\mathbf{v}) = \nabla^2 F(\mathbf{v}) - (\mathbf{v}^\top \nabla F(\mathbf{v})) \mathbf{I}.\tag{9}$$

Taking derivatives, we decompose

$$\nabla g(\mathbf{v}) = \mathcal{H}(\mathbf{v}) + \mathcal{N}(\mathbf{v}), \quad (10)$$

where the additional term $\mathcal{N}(\mathbf{v})$ is defined as

$$\mathcal{N}(\mathbf{v}) = -\mathbf{v}(\nabla F(\mathbf{v}) + \nabla^2 F(\mathbf{v})\mathbf{v})^\top. \quad (11)$$

The following lemma shows that $g(\mathbf{v})$, $\mathcal{H}(\mathbf{v})$, $\mathcal{N}(\mathbf{v})$ are Lipschitz continuous.

Lemma 5 *Given Assumption ??, we have that $g(\mathbf{v})$, $\mathcal{H}(\mathbf{v})$, $\mathcal{N}(\mathbf{v})$ are L_G, L_H, L_N -Lipschitz and $\|\mathcal{H}(\mathbf{v})\| \leq B_H$ within $\{\mathbf{v} : \|\mathbf{v}\| \leq 1, \|\mathbf{v} - \mathbf{v}^*\| \leq \delta\}$, where the constants are defined as $L_G \equiv L_K + 2L_F$, $L_H \equiv L_Q + L_F + L_K$, $L_N \equiv L_F + 3L_K + L_Q$, $B_H \equiv L_F + L_K$.*

A proof of Lemma 5 is deferred to §E.2.

For notational simplicity, we denote $\mathcal{H}_* = \mathcal{H}(\mathbf{v}^*)$ and $\mathcal{N}_* = \mathcal{N}(\mathbf{v}^*)$, and recall that \mathcal{F}_t is the filtration generated by ζ_t . Then we have the following lemma.

Lemma 6 *Under Assumptions ?? and ??, when $\eta \leq 1/(5M)$, on the event $(\|\Gamma(\mathbf{v}_{t-1}; \zeta_t)\| \leq M)$, the update rule (??) of \mathbf{v}_t can be written as*

$$\mathbf{v}_t - \mathbf{v}^* = (\mathbf{I} - \eta D\mathcal{H}_* - \eta D\mathcal{N}_*)(\mathbf{v}_{t-1} - \mathbf{v}^*) + \eta \xi_t + \eta \mathbf{R}_t + \eta^2 \mathbf{Q}_t, \quad (12)$$

where $\{\xi_t\}$ forms a vector-valued martingale difference sequence with respect to \mathcal{F}_t , ξ_t is α -sub-Weibull with parameter $G_\alpha \mathcal{V}$, \mathbf{R}_t satisfies $\|\mathbf{R}_t\| \leq (DL_H + DL_N + L_D L_G)\|\mathbf{v}_{t-1} - \mathbf{v}^*\|^2$ and \mathbf{Q}_t satisfies $\|\mathbf{Q}_t\| \leq 7M^2$.

The proof of Lemma 6 is deferred to §E.3. We define the projection of $\mathbf{v}_t - \mathbf{v}^*$ on $\mathcal{T}(\mathbf{v}^*)$ as

$$\Delta_t = (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v}_t - \mathbf{v}^*), \quad (13)$$

and the projection of \mathcal{H}_* on $\mathcal{T}(\mathbf{v}^*)$ as

$$\mathcal{M}_* = (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top})\mathcal{H}_*(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}). \quad (14)$$

Lemma 7 *Under initialization condition (??), the following properties hold:*

(i) *For all $t \geq 0$, Δ_t defined as in (13) satisfies*

$$\|(\mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v}_t - \mathbf{v}^*)\| = \frac{1}{2}\|\mathbf{v}_t - \mathbf{v}^*\|^2, \quad \|\Delta_t\|^2 = \|\mathbf{v}_t - \mathbf{v}^*\|^2 - \frac{1}{4}\|\mathbf{v}_t - \mathbf{v}^*\|^4.$$

$$\text{If } \mathbf{v}_t^\top \mathbf{v}^* \geq 0,$$

$$\|\Delta_t\|^2 \leq \|\mathbf{v}_t - \mathbf{v}^*\|^2 \leq 2\|\Delta_t\|^2. \quad (15)$$

(ii) *When $\eta \leq 1/(DB_H)$, for all $\mathbf{u} \in \mathcal{T}(\mathbf{v}^*)$,*

$$\|(\mathbf{I} - \eta D\mathcal{M}_*)^t \mathbf{u}\| \leq (1 - \eta D\mu)^t \|\mathbf{u}\|, \quad (16)$$

where \mathcal{M}_* was defined in (14).

The proof of Lemma 7 is deferred to §E.4. To interpret Lemma 7(i), we denote $\theta \equiv \angle(\mathbf{v}_t, \mathbf{v}^*) \in [0, \pi/2]$, such that $\|\mathbf{v}_t - \mathbf{v}^*\| = 2 \sin(\theta/2)$, $\Delta_t = (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v}_t - \mathbf{v}^*) = \sin \theta$, and (15) is equivalent to the trigonometric inequality

$$\sin^2 \theta = 4 \sin^2(\theta/2) \cos^2(\theta/2) \leq 4 \sin^2(\theta/2) = 2(1 - \cos \theta) \leq 2(1 - \cos \theta)(1 + \cos \theta) = 2 \sin^2 \theta.$$

By combining Lemmas 6 and 7, we have the following lemma for the update rule in terms of Δ_t :

Lemma 8 *Under Assumptions ??, ?? and initialization condition (??), when $\eta \leq 1/(5M)$, on the event $(\|\Gamma(\mathbf{v}_{t-1}; \zeta_t)\| \leq M)$, the update (??) can be written in terms of Δ_t as*

$$\Delta_t = (\mathbf{I} - \eta D\mathcal{M}_*) \Delta_{t-1} + \eta \chi_t + \eta \mathbf{S}_t + \eta^2 \mathbf{P}_t; \quad (17)$$

where $\chi_t, \mathbf{S}_t, \mathbf{P}_t \in \mathcal{T}(\mathbf{v}^*)$, $\{\chi_t\}$ forms a vector-valued martingale difference sequence with respect to \mathcal{F}_t , χ_t is α -sub-Weibull with parameter $G_\alpha \mathcal{V}$, \mathbf{S}_t satisfies $\|\mathbf{S}_t\| \leq \rho \|\mathbf{v}_{t-1} - \mathbf{v}^*\|^2$ and \mathbf{P}_t satisfies $\|\mathbf{P}_t\| \leq 7M^2$.

Proof of Lemma 8 is deferred to §E.5. Here we have $\rho = D(L_H + L_N + B_H/2) + L_D L_G$, which is consistent with its definition in (??).

Now, to analyze the iteration Δ_t we need to control its tail behavior. We define the truncated version

$$\tilde{S}_t = S_t 1_{(\mathcal{T}_M > t)}, \quad \tilde{P}_t = P_t 1_{(\mathcal{T}_M > t)}, \quad (18)$$

let $\bar{\Delta}_0 = \Delta_0$, and define the coupled process iteratively

$$\bar{\Delta}_t = (\mathbf{I} - \eta D\mathcal{M}_*) \bar{\Delta}_{t-1} + \eta \chi_t + \eta \tilde{S}_t + \eta^2 \tilde{P}_t. \quad (19)$$

The $\bar{\Delta}_t$ iteration avoids the potential issues of summation over P_t . We conclude the following lemma that characterizes the coupling relation $\bar{\Delta}_t = \Delta_t$, which allows us to analyze the coupled iteration $\bar{\Delta}_t$.

Lemma 9 *For each $t \geq 0$ we have $\bar{\Delta}_t = \Delta_t$ on the event $(\mathcal{T}_M > t)$. Furthermore, we have for all $t \geq 1$*

$$\bar{\Delta}_t = (\mathbf{I} - \eta D\mathcal{M}_*)^t \Delta_0 + \eta \sum_{s=1}^t (\mathbf{I} - \eta D\mathcal{M}_*)^{t-s} \chi_s + \eta \sum_{s=1}^t (\mathbf{I} - \eta D\mathcal{M}_*)^{t-s} \tilde{S}_s + \eta^2 \sum_{s=1}^t (\mathbf{I} - \eta D\mathcal{M}_*)^{t-s} \tilde{P}_s. \quad (20)$$

We defer the proof of Lemma 9 in §E.6.

Next we provide a lemma that tightly characterizes the approximations in (20) that $\bar{\Delta}_t \approx (\mathbf{I} - \eta D\mathcal{M}_*)^t \Delta_0$.

Lemma 10 *Let $\eta \leq \min\{1/(DB_H), 1/(5M)\}$ and $T \geq 1$. Then with probability at least*

$$1 - \left(12 + 8 \left(\frac{3}{\alpha} \right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) T \epsilon,$$

the algorithm satisfies for each $t \in [0, T]$, conditioning on $\|v_s - v^\| \leq r$ for all $s = 0, \dots, t-1$ for some $r > 0$*

$$\|\bar{\Delta}_t - (\mathbf{I} - \eta D\mathcal{M}_*)^t \Delta_0\| \leq \frac{8G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} + \frac{7\mathcal{V}^2}{D\mu} \log^{\frac{2}{\alpha}} \epsilon^{-1} \cdot \eta. \quad (21)$$

The proof of Lemma 10 is provided in §E.7.

In the following lemma we prove that when the initial iterate v_0 is sufficiently close to the minimizer v^* and r is appropriately chosen to be dependent on Δ_0 and $\tilde{\Theta}(\eta^{1/2})$, the conditioning event occurs almost surely on a high-probability event.

Lemma 11 *When initialization*

$$\|\Delta_0\| \leq \left\{ \frac{D\mu}{2^5 G_\alpha \rho}, \delta \right\},$$

for any positives η, ϵ satisfying scaling condition (??), with probability at least

$$1 - \left(14 + 8 \left(\frac{3}{\alpha} \right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) T \epsilon,$$

for all $t \in [0, T]$ we have

$$\|\Delta_t\| \leq 2 \max \left\{ \|\Delta_0\|, \frac{2^7 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} \right\},$$

and if $T_\eta^ \in [0, T]$, at time T_η^* we have*

$$\|\Delta_{T_\eta^*}\| \leq \frac{1}{2} \max \left\{ \|\Delta_0\|, \frac{2^7 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} \right\}.$$

Lemma 11, whose proof is given in §E.8, implies that the iteration keeps $\|\Delta_t\| \leq 2\|\Delta_0\|$ unless v is within a noisy neighborhood of the local minimizer v^* , where we recall the definition of Δ_t in (13).

Finally, Proposition ?? is proved by combining Lemmas 7 and 11.

B.2 PROOF OF THEOREMS ?? AND ??

In this subsection, we aim to prove Theorem ?. To deal with points with strong gradient corresponding to (i) in Definition ?, we use the following lemma that is adapted from Ge et al. [2015, Lemma 38].

Proposition 12 *Assume all conditions in Theorem ? as well as $\sqrt{2d\mathcal{V}^2 L_G D_+ \eta} < \beta$, we have on the event $(\|\nabla F(\mathbf{v}_t)\| \geq \sqrt{2d\mathcal{V}^2 L_G D_+ \eta})$ that*

$$\mathbb{E}[F(\mathbf{v}_{t+1}) - F(\mathbf{v}_t) \mid \mathcal{F}_t] \leq -0.5d\sigma^2 L_G D_-^2 \eta^2. \quad (22)$$

A core problem involves escaping from saddle points that corresponds to (iii) in Definition ?, we conclude the following modification from Ge et al. [2015, Lemma 40].

Proposition 13 *Assume all conditions in Theorem ? as well as $\sqrt{2\eta\sigma^2 L_G d D_+} < \beta$. Then on the event*

$$\left\{ \|\nabla F(\mathbf{v}_0)\| < \sqrt{2d\mathcal{V}^2 L_G D_+ \eta}, \lambda_{\min}(\mathcal{H}(\mathbf{v}_0)) \leq -\gamma \right\},$$

there is a stopping time $\mathcal{T}(\mathbf{v}_0) \leq T_{\max}$ almost surely such that

$$\mathbb{E}F(\mathbf{v}_{\mathcal{T}(\mathbf{v}_0)}) - F(\mathbf{v}_0) \leq -0.5\sigma^2 D_- \eta, \quad (23)$$

where T_{\max} is fixed and independent of \mathbf{v}_0 defined as

$$T_{\max} = 0.5\gamma^{-1} D_-^{-1} \eta^{-1} \log \left(\frac{6d\mathcal{V}}{\sigma} \right).$$

Proofs of Propositions 12 and 13 are straightforward generalization of relevant proofs of [Ge et al., 2015], and hence we omit the details.

Proof [Proof of Theorem ?] While this proof can be done in a similar fashion as Theorem 36 in Ge et al. [2015], here we provide a different proof using stopping-time techniques.

(i) Given (?), we split the state space \mathcal{S}^{d-1} into three distinct regions: let

$$\mathcal{Q}_1 = \left\{ \mathbf{v} \in \mathcal{S}^{d-1} : \|\nabla F(\mathbf{v})\| \geq \sqrt{2d\mathcal{V}^2 L_G D_+ \eta} \right\},$$

and let

$$\mathcal{Q}_2 = \left\{ \mathbf{v} \in \mathcal{S}^{d-1} : \|\nabla F(\mathbf{v})\| < \sqrt{2d\mathcal{V}^2 L_G D_+ \eta}, \lambda_{\min}(\mathcal{H}(\mathbf{v})) \leq -\gamma \right\}.$$

Define a stochastic process $\{\mathcal{T}_i\}$ s.t. $\mathcal{T}_0 = 0$, and

$$\mathcal{T}_{i+1} = \mathcal{T}_i + 1_{\mathcal{Q}_1}(\mathbf{v}_{\mathcal{T}_i}) + \mathcal{T}(\mathbf{v}_{\mathcal{T}_i}) 1_{\mathcal{Q}_2}(\mathbf{v}_{\mathcal{T}_i}), \quad (24)$$

where $\mathcal{T}(\mathbf{v}_{\mathcal{T}_i}) \leq T_{\max}$ is defined in Proposition 13. By (22) in Lemma 12 and (23) in Proposition 13, we know that on $(\mathbf{v}_{\mathcal{T}_i} \in \mathcal{Q}_1)$

$$\mathbb{E}[F(\mathbf{v}_{\mathcal{T}_{i+1}}) - F(\mathbf{v}_{\mathcal{T}_i}) \mid \mathcal{F}_{\mathcal{T}_i}] \leq -0.5d\sigma^2 L_G D_-^2 \eta^2,$$

and on $(\mathbf{v}_{\mathcal{T}_i} \in \mathcal{Q}_2)$

$$\mathbb{E}[F(\mathbf{v}_{\mathcal{T}_{i+1}}) - F(\mathbf{v}_{\mathcal{T}_i}) \mid \mathcal{F}_{\mathcal{T}_i}] \leq -0.5\sigma^2 D_- \eta.$$

Combining the above two displays and (24), we have

$$\begin{aligned} & \mathbb{E}[F(\mathbf{v}_{\mathcal{T}_{i+1}}) - F(\mathbf{v}_{\mathcal{T}_i}) \mid \mathcal{F}_{\mathcal{T}_i}] \\ & \leq -\min \left(0.5d\sigma^2 L_G D_-^2 \eta^2, \frac{0.5\sigma^2 D_- \eta}{0.5\gamma^{-1} D_-^{-1} \eta^{-1} \log \left(\frac{6d\mathcal{V}}{\sigma} \right)} \right) \cdot \mathbb{E}[\mathcal{T}_{i+1} - \mathcal{T}_i \mid \mathcal{F}_{\mathcal{T}_i}] \\ & \leq -\min \left(0.5dL_G, \gamma \log^{-1} \left(\frac{6d\mathcal{V}}{\sigma} \right) \right) \sigma^2 D_-^2 \eta^2 \cdot \mathbb{E}[\mathcal{T}_{i+1} - \mathcal{T}_i \mid \mathcal{F}_{\mathcal{T}_i}], \end{aligned} \quad (25)$$

on $\{\mathbf{v}_{\mathcal{T}_i} \in \mathcal{Q}_1 \cup \mathcal{Q}_2\}$.

- (ii) Let $\mathcal{I} \in [0, \infty]$ be the (random) first index i such that $\mathbf{v}_{\mathcal{T}_i} \in (\mathcal{Q}_1 \cup \mathcal{Q}_2)^c$. We conclude immediately that $(\mathcal{I} > i) \in \mathcal{F}_{\mathcal{T}_i}$, and $(\mathcal{I} > i) \subseteq (\mathbf{v}_{\mathcal{T}_i} \in \mathcal{Q}_1 \cup \mathcal{Q}_2)$. Applying (25) gives

$$\begin{aligned} \mathbb{E}[F(\mathbf{v}_{\mathcal{T}_\mathcal{I}}) - F(\mathbf{v}_0)] &= \mathbb{E}\left[\sum_{i=0}^{\infty} (F(\mathbf{v}_{\mathcal{T}_{i+1}}) - F(\mathbf{v}_{\mathcal{T}_i})) 1_{\mathcal{I} > i}\right] \\ &\leq -\min\left(0.5dL_G, \gamma \log^{-1}\left(\frac{6d\mathcal{V}}{\sigma}\right)\right) \sigma^2 D_-^2 \eta^2 \cdot \mathbb{E}\mathcal{T}_\mathcal{I} \\ &\leq -\min\left(0.5dL_G, \gamma \log^{-1}\left(\frac{6d\mathcal{V}}{\sigma}\right)\right) \sigma^2 D_-^2 \eta^2 \cdot T \cdot \mathbb{P}(\mathcal{T}_\mathcal{I} \geq T), \end{aligned}$$

where $T \geq 0$ is any constant. Plugging in $T = T_1$ as in (??) gives

$$\mathbb{P}(\mathcal{T}_\mathcal{I} \geq T_1) \leq \frac{\mathbb{E}[F(\mathbf{v}_0) - F(\mathbf{v}_{\mathcal{T}_\mathcal{I}})]}{\min(0.5dL_G, \gamma \log^{-1}(\frac{6d\mathcal{V}}{\sigma})) \sigma^2 D_-^2 \eta^2 \cdot T_1} \leq \frac{2\|F\|_\infty}{4\|F\|_\infty} = \frac{1}{2}.$$

In words, event $(\mathcal{T}_\mathcal{I} < T_1)$ has at least $1/2$ probability, on which the iteration \mathbf{v}_t must enter $(\mathcal{Q}_1 \cup \mathcal{Q}_2)^c$ by time T_1 at least once.

- (iii) Noting that the argument above holds for all initial points $\mathbf{v}_0 \in \mathcal{Q}_1 \cup \mathcal{Q}_2$, so one can use Markov property and conclude that within $T_1 \cdot \lceil \log_2(\kappa^{-1}) \rceil$ steps where T_1 was defined in (??), iteration $\{\mathbf{v}_t\}$ must enter $(\mathcal{Q}_1 \cup \mathcal{Q}_2)^c$ at least once with probability at least $1 - \kappa$. The rest of our proof follows from the definition of strict-saddle function. ■

Proof [Proof of Theorem ??] The conclusion is reached by directly combining Theorems ?? and ??, setting $\mathcal{A}_T = \mathcal{H}_{??}$, along with an application of strong Markov property. ■

B.3 PROOF OF PROPOSITION 1

Proof [Proof of Proposition 1] For the GEV problem setting, the gradient and the Hessian of the objective function $F(\mathbf{v})$ are

$$\begin{aligned} \nabla F(\mathbf{v}) &= -2 \frac{(\mathbf{v}^\top \mathbf{B} \mathbf{v}) \mathbf{A} \mathbf{v} - (\mathbf{v}^\top \mathbf{A} \mathbf{v}) \mathbf{B} \mathbf{v}}{(\mathbf{v}^\top \mathbf{B} \mathbf{v})^2}, \\ \nabla^2 F(\mathbf{v}) &= -2 \frac{(\mathbf{v}^\top \mathbf{B} \mathbf{v}) \mathbf{A} - (\mathbf{v}^\top \mathbf{A} \mathbf{v}) \mathbf{B} + 2(\mathbf{A} \mathbf{v} \mathbf{v}^\top \mathbf{B} - \mathbf{B} \mathbf{v} \mathbf{v}^\top \mathbf{A})}{(\mathbf{v}^\top \mathbf{B} \mathbf{v})^2} + 8 \frac{[(\mathbf{v}^\top \mathbf{B} \mathbf{v}) \mathbf{A} - (\mathbf{v}^\top \mathbf{A} \mathbf{v}) \mathbf{B}] \mathbf{v} \mathbf{v}^\top \mathbf{B}}{(\mathbf{v}^\top \mathbf{B} \mathbf{v})^3}. \end{aligned}$$

We first notice that, for $\mathbf{v} \in \{\mathbf{v} : \|\mathbf{v}\| \leq 1, \|\mathbf{v} - \mathbf{v}^*\| \leq \delta\}$,

$$\|\nabla D(\mathbf{v})\| = \|2(\mathbf{v}^\top \mathbf{B} \mathbf{v}) \mathbf{B} \mathbf{v}\| \leq 2\|\mathbf{B}\|^2,$$

which indicates that $D(\mathbf{v})$ has Lipschitz constant $L_D \equiv 2\|\mathbf{B}\|^2$. Secondly, we introduce an arbitrary unit vector \mathbf{w} and take derivative of vector $\nabla^2 F(\mathbf{v}) \mathbf{w}$ w.r.t. \mathbf{v} as

$$\begin{aligned} \nabla_{\mathbf{v}} [\nabla^2 F(\mathbf{v}) \mathbf{w}] &= -2 \frac{2\mathbf{A} \mathbf{w} \mathbf{v}^\top \mathbf{B} - 2\mathbf{B} \mathbf{v} \mathbf{v}^\top \mathbf{A} + 2(\mathbf{v}^\top \mathbf{B} \mathbf{w}) \mathbf{A} + 2\mathbf{A} \mathbf{v} \mathbf{w}^\top \mathbf{B} - 2(\mathbf{v}^\top \mathbf{A} \mathbf{w}) \mathbf{B} - 2\mathbf{B} \mathbf{v} \mathbf{w}^\top \mathbf{A}}{(\mathbf{v}^\top \mathbf{B} \mathbf{v})^2} \\ &\quad + 8 \frac{[(\mathbf{v}^\top \mathbf{B} \mathbf{v}) \mathbf{A} - (\mathbf{v}^\top \mathbf{A} \mathbf{v}) \mathbf{B} + 2(\mathbf{A} \mathbf{v} \mathbf{v}^\top \mathbf{B} - \mathbf{B} \mathbf{v} \mathbf{v}^\top \mathbf{A})] \mathbf{w} \mathbf{v}^\top \mathbf{B}}{(\mathbf{v}^\top \mathbf{B} \mathbf{v})^3} \\ &\quad + 8 \frac{[(\mathbf{v}^\top \mathbf{B} \mathbf{v}) \mathbf{A} - (\mathbf{v}^\top \mathbf{A} \mathbf{v}) \mathbf{B}] \mathbf{v} \mathbf{w}^\top \mathbf{B}}{(\mathbf{v}^\top \mathbf{B} \mathbf{v})^3} \\ &\quad + 8 \frac{(\mathbf{v}^\top \mathbf{B} \mathbf{w}) [(\mathbf{v}^\top \mathbf{B} \mathbf{v}) \mathbf{A} - (\mathbf{v}^\top \mathbf{A} \mathbf{v}) \mathbf{B} + 2(\mathbf{A} \mathbf{v} \mathbf{v}^\top \mathbf{B} - \mathbf{B} \mathbf{v} \mathbf{v}^\top \mathbf{A})]}{(\mathbf{v}^\top \mathbf{B} \mathbf{v})^3} \\ &\quad - 48 \left[\frac{(\mathbf{v}^\top \mathbf{B} \mathbf{w}) [(\mathbf{v}^\top \mathbf{B} \mathbf{v}) \mathbf{A} - (\mathbf{v}^\top \mathbf{A} \mathbf{v}) \mathbf{B}] \mathbf{v} \mathbf{v}^\top \mathbf{B}}{(\mathbf{v}^\top \mathbf{B} \mathbf{v})^4} \right]. \end{aligned}$$

The five terms on the right-hand side have norm bounded by $\frac{24\|\mathbf{A}\|\|\mathbf{B}\|}{(1-\delta)^2\lambda_{\min}^2(\mathbf{B})}$, $\frac{48\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}$, $\frac{16\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}$, $\frac{48\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}$, $\frac{96\|\mathbf{A}\|\|\mathbf{B}\|^3}{(1-\delta)^4\lambda_{\min}^4(\mathbf{B})}$ respectively, which implies that

$$\|\nabla_{\mathbf{v}} [\nabla^2 F(\mathbf{v})\mathbf{w}]\| \leq \frac{232\|\mathbf{A}\|\|\mathbf{B}\|^3}{\lambda_{\min}^4(\mathbf{B})}.$$

Therefore, for all $\mathbf{v}_1, \mathbf{v}_2 \in \{\mathbf{v} : \|\mathbf{v}\| \leq 1, \|\mathbf{v} - \mathbf{v}^*\| \leq \delta\}$, we have

$$\|\nabla^2 F(\mathbf{v}_1) - \nabla^2 F(\mathbf{v}_2)\| = \max_{\|\mathbf{w}\|=1} \|\nabla^2 F(\mathbf{v}_1)\mathbf{w} - \nabla^2 F(\mathbf{v}_2)\mathbf{w}\| \leq \frac{232\|\mathbf{A}\|\|\mathbf{B}\|^3}{(1-\delta)^4\lambda_{\min}^4(\mathbf{B})} \|\mathbf{v}_1 - \mathbf{v}_2\|,$$

indicating $\nabla^2 F(\mathbf{v})$ has Lipschitz constant $L_Q \equiv \frac{232\|\mathbf{A}\|\|\mathbf{B}\|^3}{(1-\delta)^4\lambda_{\min}^4(\mathbf{B})}$.

Similarly, we also notice for all $\mathbf{v} \in \{\mathbf{v} : \|\mathbf{v}\| \leq 1, \|\mathbf{v} - \mathbf{v}^*\| \leq \delta\}$,

$$\|\nabla F(\mathbf{v})\| \leq \frac{4\|\mathbf{A}\|\|\mathbf{B}\|}{(1-\delta)^2\lambda_{\min}^2(\mathbf{B})}, \quad \|\nabla^2 F(\mathbf{v})\| \leq \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})},$$

which indicates that $F(\mathbf{v})$ has Lipschitz constant $L_F \equiv \frac{4\|\mathbf{A}\|\|\mathbf{B}\|}{(1-\delta)^2\lambda_{\min}^2(\mathbf{B})}$ and $\nabla F(\mathbf{v})$ has Lipschitz constant $L_K \equiv \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}$. ■

B.4 PROOF OF THEOREM ??

To prove Theorem ??, we first present the following Lemma 14 on a linear representation of $\mathcal{M}_*(\bar{\mathbf{v}}_T^{(\eta)} - \mathbf{v}^*)$.

Lemma 14 (Representation Lemma) *Under Assumptions ??, ?? and given initialization condition (??), for any $T \geq K_{\eta,\epsilon}T_{\eta}^*$ and positive constants η, ϵ satisfying the scaling condition*

$$5\mathcal{V} \log^{1/\alpha} \epsilon^{-1} \cdot \eta \leq 1,$$

we have

$$\begin{aligned} \mathcal{M}_* \left(\bar{\mathbf{v}}_T^{(\eta)} - \mathbf{v}^* \right) &= \frac{1}{D(T - K_{\eta,\epsilon}T_{\eta}^*)} \sum_{t=K_{\eta,\epsilon}T_{\eta}^*+1}^T \boldsymbol{\chi}_{t+1} + \frac{1}{D(T - K_{\eta,\epsilon}T_{\eta}^*)} \sum_{t=K_{\eta,\epsilon}T_{\eta}^*+1}^T \mathbf{S}_{t+1} \\ &\quad + \frac{\eta}{D(T - K_{\eta,\epsilon}T_{\eta}^*)} \sum_{t=K_{\eta,\epsilon}T_{\eta}^*+1}^T \mathbf{P}_{t+1} + \frac{1}{D(T - K_{\eta,\epsilon}T_{\eta}^*)\eta} (\boldsymbol{\Delta}_{K_{\eta,\epsilon}T_{\eta}^*+1} - \boldsymbol{\Delta}_{T+1}), \end{aligned} \quad (26)$$

where $\boldsymbol{\chi}_t, \mathbf{S}_t, \mathbf{P}_t$ are vectors in the tangent space $\mathcal{T}(\mathbf{v}^*)$. Here $\boldsymbol{\chi}_t$ is defined as

$$\boldsymbol{\chi}_t \equiv (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top})(\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t) - D(\mathbf{v}_{t-1})\nabla F(\mathbf{v}_{t-1})), \quad (27)$$

which is α -sub-Weibull with parameter $G_{\alpha}\mathcal{V}$. The sequence $\{\boldsymbol{\chi}_t\}$ forms a vector-valued martingale difference sequence with respect to \mathcal{F}_t . \mathbf{S}_t satisfies $\|\mathbf{S}_t\| \leq \rho\|\mathbf{v}_{t-1} - \mathbf{v}^*\|^2$. On the event $\mathcal{H}_{??}$ defined in Theorem ??, using a total sample size $T+1$, each \mathbf{P}_t satisfies $\|\mathbf{P}_t\| \leq 7\mathcal{V}^2 \log^{2/\alpha} \epsilon^{-1}$.

Proof [Proof of Lemma 14] Telescoping (17) in Lemma 8 for $t = K_{\eta,\epsilon}T_{\eta}^* + 2, \dots, T+1$ gives

$$\begin{aligned} \eta D \mathcal{M}_* \sum_{t=K_{\eta,\epsilon}T_{\eta}^*+1}^T \boldsymbol{\Delta}_t &= (\boldsymbol{\Delta}_{K_{\eta,\epsilon}T_{\eta}^*+1} - \boldsymbol{\Delta}_{T+1}) + \eta \sum_{t=K_{\eta,\epsilon}T_{\eta}^*+1}^T \boldsymbol{\chi}_{t+1} \\ &\quad + \eta \sum_{t=K_{\eta,\epsilon}T_{\eta}^*+1}^T \mathbf{S}_{t+1} + \eta^2 \sum_{t=K_{\eta,\epsilon}T_{\eta}^*+1}^T \mathbf{P}_{t+1}. \end{aligned}$$

Plugging in the definitions of $\Delta_t, \bar{v}_T^{(\eta)}$ in (13), (??) gives (26). For event $\mathcal{H}_{??}$ defined in Theorem ?? using total sample size $T + 1$, the proof of Lemma 11 in §E.8 shows that $\mathcal{H}_{??} \subseteq \{\|\Gamma(\mathbf{v}_{t-1}; \zeta_t)\| \leq M : 1 \leq t \leq T + 2\}$. The rest of Lemma 14 directly follows Lemma 8. \blacksquare

With Lemma 14 in hand, we are ready to prove Theorem ??.

Proof [Proof of Theorem ??] For a given T , we apply Theorem ?? and Lemma 14 with $\epsilon = 1/T^2$, such that $\mathbb{P}(\mathcal{H}_{??}) \rightarrow 1$ and the scaling condition (??) is satisfied under condition (??). Using a coupling approach we can safely ignore the small probability event and concentrate on the event $\mathcal{H}_{??}$, where we have

$$\left\| \frac{1}{D(T - K_{\eta, \epsilon} T_{\eta}^*)} \sum_{t=K_{\eta, \epsilon} T_{\eta}^* + 1}^T \mathbf{S}_{t+1} \right\| \leq \frac{2^{\frac{\alpha+2}{2}+17} \rho G_{\alpha}^2 \mathcal{V}^2}{D^2 \mu} \eta \log^{\frac{\alpha+2}{\alpha}} T,$$

$$\left\| \frac{\eta}{D(T - K_{\eta, \epsilon} T_{\eta}^*)} \sum_{t=K_{\eta, \epsilon} T_{\eta}^* + 1}^T \mathbf{P}_{t+1} \right\| \leq \frac{7 \cdot 2^{\frac{2}{\alpha}} \mathcal{V}^2}{D} \eta \log^{\frac{2}{\alpha}} T.$$

Using the relation $\|\Delta_t\| \leq \|\mathbf{v}_t - \mathbf{v}^*\| \leq \sqrt{2}\|\Delta_t\|$, given in Proposition ??, and applying Theorem ?? on event $\mathcal{H}_{??}$ we also have

$$\left\| \frac{1}{D(T - K_{\eta, \epsilon} T_{\eta}^*) \eta} (\Delta_{K_{\eta, \epsilon} T_{\eta}^* + 1} - \Delta_{T+1}) \right\| \leq \frac{2^{\frac{\alpha+2}{2}+\frac{17}{2}+1} G_{\alpha} \mathcal{V}}{\sqrt{D^3 \mu}} \frac{\log^{\frac{\alpha+2}{2\alpha}} T}{(T - K_{\eta, \epsilon} T_{\eta}^*) \eta^{1/2}}.$$

Under condition (??), as $T \rightarrow \infty, \eta \rightarrow 0$, we have the following almost-sure convergences

$$\frac{\sqrt{T}}{D(T - K_{\eta, \epsilon} T_{\eta}^*)} \sum_{t=K_{\eta, \epsilon} T_{\eta}^* + 1}^T \mathbf{S}_{t+1} \rightarrow \mathbf{0} \quad \text{a.s.}$$

$$\frac{\eta \sqrt{T}}{D(T - K_{\eta, \epsilon} T_{\eta}^*)} \sum_{t=K_{\eta, \epsilon} T_{\eta}^* + 1}^T \mathbf{P}_{t+1} \rightarrow \mathbf{0} \quad \text{a.s.}$$

$$\frac{\sqrt{T}}{D(T - K_{\eta, \epsilon} T_{\eta}^*) \eta} (\Delta_{K_{\eta, \epsilon} T_{\eta}^* + 1} - \Delta_{T+1}) \rightarrow \mathbf{0} \quad \text{a.s.}$$

From (??) and (27), the covariance matrix of ξ_t —i.e., the projection of scaled-gradient noise onto the tangent space $\mathcal{T}(\mathbf{v}^*)$ —can be denoted by

$$\Phi(\mathbf{v}_{t-1}) \equiv (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \Sigma(\mathbf{v}_{t-1}) (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}).$$

We denote the covariance matrix at local minimizer \mathbf{v}^* as $\Phi_* \equiv \Phi(\mathbf{v}^*) = (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \Sigma_*(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top})$. Using the central limit theorem and the Slutsky theorem, we have the following convergence-in-distribution result under the condition (??) as $T \rightarrow \infty, \eta \rightarrow 0$:

$$\frac{1}{\sqrt{T}} \sum_{t=K_{\eta, \epsilon} T_{\eta}^* + 1}^T \chi_{t+1} \xrightarrow{d} N(\mathbf{0}, \Phi_*).$$

Combining these results with (26) in Lemma 14, under condition (??), as $T \rightarrow \infty, \eta \rightarrow 0$ we have convergence in distribution:

$$\sqrt{T} \mathcal{M}_* \left(\bar{\mathbf{v}}_T^{(\eta)} - \mathbf{v}^* \right) \xrightarrow{d} N(\mathbf{0}, D^{-2} \cdot \Phi_*). \quad (28)$$

Since $\mathcal{M}_*^+ \mathcal{M}_* = \mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}$ and $\mathcal{M}_*^+ \Phi_* \mathcal{M}_*^+ = \mathcal{M}_*^+ \Sigma_* \mathcal{M}_*^+$, (28) is equivalent to

$$\sqrt{T} (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \left(\bar{\mathbf{v}}_T^{(\eta)} - \mathbf{v}^* \right) \xrightarrow{d} N(\mathbf{0}, D^{-2} \cdot \mathcal{M}_*^+ \Sigma_* \mathcal{M}_*^+), \quad (29)$$

which omits the asymptotic analysis in the direction parallel to \mathbf{v}^* . To study the asymptotic property of $\mathbf{v}^* \mathbf{v}^{*\top} (\bar{\mathbf{v}}_T^{(\eta)} - \mathbf{v}^*)$, we first notice that in Lemma 7 in §B.1 we know that for all $\mathbf{v} \in \mathbb{R}^d$ with $\|\mathbf{v}\| = 1$, $\|\mathbf{v}^* \mathbf{v}^{*\top} (\mathbf{v} - \mathbf{v}^*)\| = 1 - \mathbf{v}^{*\top} \mathbf{v} = \frac{1}{2} \|\mathbf{v} - \mathbf{v}^*\|^2$.

Applying Theorem ??, on event $\mathcal{H}_{??}$ we have:

$$\begin{aligned} \left\| \sqrt{T} \cdot \mathbf{v}^* \mathbf{v}^{*\top} (\bar{\mathbf{v}}_T^{(\eta)} - \mathbf{v}^*) \right\| &= \frac{1}{2\sqrt{T}} \sum_{t=K_{\eta,\epsilon}T_{\eta}^*+1}^T \|\mathbf{v}_t - \mathbf{v}^*\|^2 \\ &\leq \frac{2^{\frac{\alpha+2}{\alpha}+17} G_{\alpha}^2 \mathcal{V}^2}{D\mu} \cdot \frac{\eta(T - K_{\eta,\epsilon}T_{\eta}^*) \log^{\frac{\alpha+2}{\alpha}} T}{\sqrt{T}} \lesssim \sqrt{\eta^2 T \log^{\frac{2\alpha+4}{\alpha}} T} \rightarrow 0, \end{aligned}$$

where in the second line we used the first condition in (??). Under condition (??), as $T \rightarrow \infty, \eta \rightarrow 0$, we have almost-sure convergence

$$\sqrt{T} \cdot \mathbf{v}^* \mathbf{v}^{*\top} (\bar{\mathbf{v}}_T^{(\eta)} - \mathbf{v}^*) \rightarrow \mathbf{0} \quad \text{a.s.} \quad (30)$$

Adding up (29) and (30) and applying the Slutsky theorem, we conclude (??) and Theorem ??. \blacksquare

B.5 PROOF OF PROPOSITION ??

Proof [Proof of Proposition ??] For notational simplicity, we denote vector $\mathbf{v} \in \mathbb{R}^{d_x+d_y}$ as $\mathbf{v}^\top = (\mathbf{v}_x^\top, \mathbf{v}_y^\top)$ for $\mathbf{v}_x \in \mathbb{R}^{d_x}, \mathbf{v}_y \in \mathbb{R}^{d_y}$. For any vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^{d_x}$ with $\|\mathbf{w}_1\| \leq 1, \|\mathbf{w}_2\| \leq 1$, using Lemma 17 we have

$$\|\mathbf{w}_1^\top \mathbf{X} \mathbf{X}^\top \mathbf{w}_2\|_{\psi_1} \leq \|\mathbf{w}_1^\top \mathbf{X}\|_{\psi_2} \|\mathbf{w}_2^\top \mathbf{X}\|_{\psi_2} \leq \mathcal{V}_x^2,$$

which indicates that

$$\|\mathbf{v}_x^\top \mathbf{X} \mathbf{X}^\top \mathbf{v}_x\|_{\psi_1} \leq \mathcal{V}_x^2, \quad \|\mathbf{X} \mathbf{X}^\top \mathbf{v}_x\|_{\psi_1} \leq \mathcal{V}_x^2.$$

Similarly, we can show

$$\|\mathbf{v}_y^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{v}_y\|_{\psi_1} \leq \mathcal{V}_y^2, \quad \|\mathbf{Y} \mathbf{Y}^\top \mathbf{v}_y\|_{\psi_1} \leq \mathcal{V}_y^2,$$

and

$$\|\mathbf{v}_x^\top \mathbf{X} \mathbf{Y}^\top \mathbf{v}_y\|_{\psi_1} \leq \mathcal{V}_x \mathcal{V}_y, \quad \|\mathbf{X} \mathbf{Y}^\top \mathbf{v}_y\|_{\psi_1} \leq \mathcal{V}_x \mathcal{V}_y, \quad \|\mathbf{Y} \mathbf{X}^\top \mathbf{v}_x\|_{\psi_1} \leq \mathcal{V}_x \mathcal{V}_y.$$

Combining all above inequalities and using Lemma 16 yields

$$\left\| \mathbf{v}^\top \tilde{\mathbf{A}} \mathbf{v} \right\|_{\psi_1} \leq 2\mathcal{V}_x \mathcal{V}_y, \quad \left\| \tilde{\mathbf{A}} \mathbf{v} \right\|_{\psi_1} \leq 2\mathcal{V}_x \mathcal{V}_y, \quad \left\| \mathbf{v}^\top \tilde{\mathbf{B}}' \mathbf{v} \right\|_{\psi_1} \leq \mathcal{V}_x^2 + \mathcal{V}_y^2, \quad \left\| \tilde{\mathbf{B}}' \mathbf{v} \right\|_{\psi_1} \leq \mathcal{V}_x^2 + \mathcal{V}_y^2.$$

By applying Lemmas 17 and 18, in CCA problem we have stochastic scaled-gradient satisfying

$$\begin{aligned} \left\| (\mathbf{v}^\top \tilde{\mathbf{B}}' \mathbf{v}) \tilde{\mathbf{A}} \mathbf{v} - (\mathbf{v}^\top \tilde{\mathbf{A}} \mathbf{v}) \tilde{\mathbf{B}}' \mathbf{v} \right\|_{\psi_{1/2}} &\leq G_{1/2} \left(\left\| \mathbf{v}^\top \tilde{\mathbf{B}}' \mathbf{v} \right\|_{\psi_1} \left\| \tilde{\mathbf{A}} \mathbf{v} \right\|_{\psi_1} + \left\| \mathbf{v}^\top \tilde{\mathbf{A}} \mathbf{v} \right\|_{\psi_1} \left\| \tilde{\mathbf{B}}' \mathbf{v} \right\|_{\psi_1} \right) \\ &\leq 400(\mathcal{V}_x^2 + \mathcal{V}_y^2) \mathcal{V}_x \mathcal{V}_y. \end{aligned}$$

Hence Assumption ?? holds for $\mathcal{V} = 400(\mathcal{V}_x^2 + \mathcal{V}_y^2) \mathcal{V}_x \mathcal{V}_y$ and $\alpha = 1/2$. \blacksquare

C PRELIMINARIES FOR ORLICZ- ψ_α NORM

Of similar style as [Li and Jordan, 2021, §E] we collect in this section some facts for Orlicz- ψ_α norm for our usage. We start with its definition:

Definition 15 (Orlicz ψ_α -norm) For a continuous, monotonically increasing and convex function $\psi(x)$ defined for all $x > 0$ satisfying $\psi(0) = 0$ and $\lim_{x \rightarrow \infty} \psi(x) = \infty$, we define the Orlicz ψ -norm for a random variable X as

$$\|X\|_\psi \equiv \inf \left\{ K > 0 : \mathbb{E} \psi \left(\frac{|X|}{K} \right) \leq 1 \right\}.$$

As a commonly used special case, we consider function $\psi_\alpha(x) \equiv \exp(x^\alpha) - 1$ and define the Orlicz ψ_α -norm for a random variable X as

$$\|X\|_{\psi_\alpha} \equiv \inf \left\{ K > 0 : \mathbb{E} \exp \left(\frac{|X|^\alpha}{K^\alpha} \right) \leq 2 \right\}.$$

Lemma 16 When $\psi(x)$ is monotonically increasing and convex for $x > 0$, for any random variables X, Y with finite Orlicz ψ -norm, the triangle inequality holds

$$\|X + Y\|_{\psi} \leq \|X\|_{\psi} + \|Y\|_{\psi}.$$

For all $\alpha \geq 1$, the above inequality holds when $\|\cdot\|_{\psi}$ is taken as the Orlicz ψ_{α} -norm.

Proof [Proof of Lemma 16] Let K_1, K_2 denote the Orlicz ψ -norms of X and Y . Because $\psi(x)$ is monotonically increasing and convex, we have

$$\begin{aligned} \psi\left(\frac{|X+Y|}{K_1+K_2}\right) &\leq \psi\left(\frac{K_1}{K_1+K_2} \cdot \frac{|X|}{K_1} + \frac{K_2}{K_1+K_2} \cdot \frac{|Y|}{K_2}\right) \\ &\leq \frac{K_1}{K_1+K_2} \cdot \psi\left(\frac{|X|}{K_1}\right) + \frac{K_2}{K_1+K_2} \cdot \psi\left(\frac{|Y|}{K_2}\right), \end{aligned}$$

which implies

$$\mathbb{E}\psi\left(\frac{|X+Y|}{K_1+K_2}\right) \leq 1, \quad \text{i.e. } \|X+Y\|_{\psi} \leq \|X\|_{\psi} + \|Y\|_{\psi},$$

yielding the lemma. ■

Lemma 17 Let X and Y be random variables with finite ψ_{α} -norm for some $\alpha \geq 1$, then

$$\|XY\|_{\psi_{\alpha/2}} \leq \|X\|_{\psi_{\alpha}} \|Y\|_{\psi_{\alpha}}.$$

Proof [Proof of Lemma 17] Denote $A \equiv X/\|X\|_{\psi_{\alpha}}$, $B \equiv Y/\|Y\|_{\psi_{\alpha}}$, then $\|A\|_{\psi_{\alpha}} = \|B\|_{\psi_{\alpha}} = 1$. Using the elementary inequality

$$|AB| \leq \frac{1}{4}(|A| + |B|)^2,$$

and the triangle inequality in Lemma 16 we have that

$$\|AB\|_{\psi_{\alpha/2}} \leq \frac{1}{4} \|(|A| + |B|)^2\|_{\psi_{\alpha/2}} = \frac{1}{4} \| |A| + |B| \|_{\psi_{\alpha}}^2 \leq \frac{1}{4} (\|A\|_{\psi_{\alpha}} + \|B\|_{\psi_{\alpha}})^2 = 1.$$

Multiplying both sides of the inequality by $\|X\|_{\psi_{\alpha}} \|Y\|_{\psi_{\alpha}}$ gives the desired result. ■

Lemma 18 For any random variables X, Y with finite Orlicz ψ_{α} -norm, the following inequalities hold

$$\|X + Y\|_{\psi_{\alpha}} \leq \log_2^{1/\alpha}(1 + e^{1/\alpha})(\|X\|_{\psi_{\alpha}} + \|Y\|_{\psi_{\alpha}}), \quad \|\mathbb{E}X\|_{\psi_{\alpha}} \leq \log_2^{1/\alpha}(1 + e^{1/\alpha})\|X\|_{\psi_{\alpha}},$$

and

$$\|X - \mathbb{E}X\|_{\psi_{\alpha}} \leq \log_2^{1/\alpha}(1 + e^{1/\alpha}) \left(1 + \log_2^{1/\alpha}(1 + e^{1/\alpha})\right) \|X\|_{\psi_{\alpha}}.$$

Proof [Proof of Lemma 18] Recall that when $\alpha \in (0, 1)$, $\psi_{\alpha}(x)$ does *not* satisfy convexity when x is around 0. Let $\tilde{\psi}_{\alpha}(x)$ be

$$\tilde{\psi}_{\alpha}(x) = \begin{cases} \exp(x^{\alpha}) - 1 & x \geq x_* \\ \frac{x}{x_*} (\exp(x_*^{\alpha}) - 1) & x \in [0, x_*) \end{cases}.$$

for some appropriate $x_* > 0$, so as to make the function convex. Here x_* is chosen such that the tangent line of function ψ_{α} at x_* passes through origin, i.e.

$$\psi'_{\alpha}(x_*) = \alpha x_*^{\alpha-1} \exp(x_*^{\alpha}) = \frac{\exp(x_*^{\alpha}) - 1}{x_*} = \tilde{\psi}'_{\alpha}(x_*).$$

Simplifying it gives us a transcendental equation

$$(1 - \alpha x_*^{\alpha}) \exp(x_*^{\alpha}) = 1.$$

We easily find that $x_*^\alpha \leq 1/\alpha$. Because $\psi_\alpha(x)$ is concave on $(0, (\frac{1}{\alpha} - 1)^{1/\alpha})$ and convex on $((\frac{1}{\alpha} - 1)^{1/\alpha}, \infty)$, we have $\psi_\alpha(x) \geq \tilde{\psi}_\alpha(x) \geq 0$ for all $x \geq 0$, and hence

$$0 \leq \psi_\alpha(x) - \tilde{\psi}_\alpha(x) \leq \psi_\alpha(x_*) \leq e^{1/\alpha} - 1. \quad (31)$$

Let K_1, K_2 denote the Orlicz ψ_α -norms of X and Y , then

$$\mathbb{E}\tilde{\psi}_\alpha\left(\frac{|X|}{K_1}\right) \leq \mathbb{E}\psi_\alpha\left(\frac{|X|}{K_1}\right) \leq 1, \quad \mathbb{E}\tilde{\psi}_\alpha\left(\frac{|Y|}{K_2}\right) \leq \mathbb{E}\psi_\alpha\left(\frac{|Y|}{K_2}\right) \leq 1.$$

By applying the triangle inequality in Lemma 16 and using (31), we have

$$\begin{aligned} \mathbb{E}\psi_\alpha\left(\frac{|X+Y|}{K_1+K_2}\right) &\leq \mathbb{E}\tilde{\psi}_\alpha\left(\frac{|X+Y|}{K_1+K_2}\right) + e^{1/\alpha} - 1 \leq e^{1/\alpha}, \\ \mathbb{E}\psi_\alpha\left(\frac{|\mathbb{E}X|}{K_1}\right) &\leq \mathbb{E}\tilde{\psi}_\alpha\left(\frac{|\mathbb{E}X|}{K_1}\right) + e^{1/\alpha} - 1 \leq e^{1/\alpha}. \end{aligned}$$

By applying Jensen's inequality to concave function $J_\alpha(z) = z^{\log_{1+e^{1/\alpha}} 2}$, we have

$$\begin{aligned} \mathbb{E}\psi_\alpha\left(\frac{|X+Y|}{\log_2^{1/\alpha}(1+e^{1/\alpha})(K_1+K_2)}\right) &= \mathbb{E}J_\alpha\left(\exp\left(\frac{|X+Y|^\alpha}{(K_1+K_2)^\alpha}\right)\right) - 1 \\ &\leq J_\alpha\left(\mathbb{E}\exp\left(\frac{|X+Y|^\alpha}{(K_1+K_2)^\alpha}\right)\right) - 1 \leq 1, \end{aligned}$$

and

$$\mathbb{E}\psi_\alpha\left(\frac{|\mathbb{E}X|}{\log_2^{1/\alpha}(1+e^{1/\alpha})K_1}\right) = \mathbb{E}J_\alpha\left(\exp\left(\frac{|\mathbb{E}X|^\alpha}{K_1^\alpha}\right)\right) - 1 \leq J_\alpha\left(\mathbb{E}\exp\left(\frac{|\mathbb{E}X|^\alpha}{K_1^\alpha}\right)\right) - 1 \leq 1,$$

which implies

$$\|X+Y\|_{\psi_\alpha} \leq \log_2^{1/\alpha}(1+e^{1/\alpha})(\|X\|_{\psi_\alpha} + \|Y\|_{\psi_\alpha}), \quad \|\mathbb{E}X\|_{\psi_\alpha} \leq \log_2^{1/\alpha}(1+e^{1/\alpha})\|X\|_{\psi_\alpha},$$

and

$$\|X - \mathbb{E}X\|_{\psi_\alpha} \leq \log_2^{1/\alpha}(1+e^{1/\alpha})(\|X\|_{\psi_\alpha} + \|\mathbb{E}X\|_{\psi_\alpha}) \leq \log_2^{1/\alpha}(1+e^{1/\alpha})\left(1 + \log_2^{1/\alpha}(1+e^{1/\alpha})\right)\|X\|_{\psi_\alpha}.$$

■

Now we proceed with the definition of Orlicz ψ_α -norm for random vectors.

Definition 19 For a random vector $\mathbf{X} \in \mathbb{R}^d$, its Orlicz ψ_α -norm is defined as

$$\|\mathbf{X}\|_{\psi_\alpha} \equiv \inf \left\{ K > 0 : \mathbb{E} \exp \left(\frac{\|\mathbf{X}\|^\alpha}{K^\alpha} \right) \leq 2 \right\}.$$

Seeing the above definition, a random vector \mathbf{X} is called *sub-Gaussian* if $\|\mathbf{X}\|_{\psi_2} < \infty$, and is called *sub-Exponential* if $\|\mathbf{X}\|_{\psi_1} < \infty$.

Remark 20 We notice that $\|\mathbf{X}\|_{\psi_\alpha}$ equals to the Orlicz ψ_α -norm of random variable (scalar) $\|\mathbf{X}\|$. Using this relation, we can easily extend all above results of random variables to random vectors with the same positive factors and dependency on α .

D ESTIMATION OF THE STRICT-SADDLE PARAMETERS

The goal of this section is to detail the proof of Lemma 3 that estimates the strict-saddle parameters. We first compute the manifold gradient and Hessian in the following Lemma 21:

Lemma 21 *The manifold gradient and Hessian can be computed as*

$$g(\mathbf{x}) = -2 \frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} \mathbf{x}, \quad (32)$$

$$\mathcal{H}(\mathbf{x}) = -2 \frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B} + 2(\mathbf{A} \mathbf{x} \mathbf{x}^\top \mathbf{B} - \mathbf{B} \mathbf{x} \mathbf{x}^\top \mathbf{A})}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} + 8 \frac{[(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B}] \mathbf{x} \mathbf{x}^\top \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^3}. \quad (33)$$

Proof The constrained optimization problem has $c(\mathbf{x}) = \|\mathbf{x}\|^2 - 1$ so the Lagrangian is

$$\mathcal{L}(\mathbf{x}; \mu) = -\frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{B} \mathbf{x}} - \mu(\mathbf{x}^\top \mathbf{x} - 1).$$

According to the constrained optimization theory in Nocedal and Wright [2006], since (i) there is one constraint (ii) the gradient $g(\mathbf{x}) = 2\mathbf{x}$ on constraint has constant norm 2, it satisfies some 2-RLICQ condition. The feasible value of Lagrangian multiplier $\mu^*(\mathbf{x})$ has

$$\mu^*(\mathbf{x}) = \arg \min_{\mu} \|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mu)\|^2.$$

Let

$$\Lambda(\mathbf{x}) = \frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2}.$$

Then we have

$$\nabla \mathcal{L}(\mathbf{x}; \mu) = -2\Lambda(\mathbf{x})\mathbf{x} - 2\mu\mathbf{x},$$

and hence

$$\begin{aligned} \|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}; \mu)\|^2 &= 4 \|\Lambda(\mathbf{x})\mathbf{x} + \mu\mathbf{x}\|^2 = 4 \|\Lambda(\mathbf{x})\mathbf{x} + \mu\mathbf{x}\|^2 \\ &= 4 (\mathbf{x}^\top \Lambda(\mathbf{x}) \Lambda(\mathbf{x}) \mathbf{x} + 2(\mathbf{x}^\top \Lambda(\mathbf{x}) \mathbf{x}) \mu + (\mathbf{x}^\top \mathbf{x}) \mu^2). \end{aligned}$$

Solving this problem gives $\mu^*(\mathbf{x})$ for $\mathbf{x} \in \mathcal{S}^{d-1}$:

$$\mu^*(\mathbf{x}) = -\mathbf{x}^\top \Lambda(\mathbf{x}) \mathbf{x} = -\frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{x}^\top \mathbf{A} \mathbf{x} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{x}^\top \mathbf{B} \mathbf{x}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} = 0.$$

The manifold gradient can hence be computed as

$$g(\mathbf{x}) = \nabla L(\mathbf{x}; \mu) \big|_{\mu=\mu^*(\mathbf{x})} = -2\Lambda(\mathbf{x})\mathbf{x} - 2\mu^*(\mathbf{x})\mathbf{x} = -2 \frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} \mathbf{x},$$

concluding (32). For manifold Hessian, we can compute it as

$$\begin{aligned} \mathcal{H}(\mathbf{x}) &= \nabla^2 L(\mathbf{x}; \mu) \big|_{\mu=\mu^*(\mathbf{x})} = -2 \nabla \left[\frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} \mathbf{x} \right] \\ &= -2 \frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B} + 2(\mathbf{A} \mathbf{x} \mathbf{x}^\top \mathbf{B} - \mathbf{B} \mathbf{x} \mathbf{x}^\top \mathbf{A})}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} + 4 \frac{[(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B}] \mathbf{x} \mathbf{x}^\top (2\mathbf{B})}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^3}. \end{aligned}$$

This proves (33) and concludes the lemma. ■

We prove the Hessian smoothness and give the Lipschitz constant for both manifold gradient and Hessian, as in the following lemmas.

Lemma 22 *There are Lipschitz constants*

$$L_G \equiv \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{\lambda_{\min}^3(\mathbf{B})}, \quad L_H \equiv \frac{56\|\mathbf{A}\|\|\mathbf{B}\|^3}{\lambda_{\min}^4(\mathbf{B})},$$

such that for all $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{S}^{d-1}$ we have

$$\|\mathcal{H}(\mathbf{z})\| \leq L_G, \quad (34)$$

and

$$\|\mathcal{H}(\mathbf{z}_1) - \mathcal{H}(\mathbf{z}_2)\| \leq L_H \|\mathbf{z}_1 - \mathbf{z}_2\|. \quad (35)$$

In addition, we have from above two

$$\left\| P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z}')}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')} \right\| \leq (2L_G + L_H) \|\mathbf{z} - \mathbf{z}'\|. \quad (36)$$

In fact, in this lemma one can replace $\|\mathbf{A}\|$ by the norm $\|\mathbf{A} - c\mathbf{B}\|$ for any constant scalar c .

Proof [Proof of Lemma 22] Note

$$\|g(\mathbf{x})\| \leq \frac{\|\mathbf{B}\|}{\lambda_{\min}^2(\mathbf{B})} \|\mathbf{A}\|,$$

and

$$\|\mathcal{H}(\mathbf{x})\| \leq 2 \frac{2\|\mathbf{B}\|\|\mathbf{A}\| + 4\|\mathbf{A}\|\|\mathbf{B}\|}{\lambda_{\min}^2(\mathbf{B})} + 8 \frac{2\|\mathbf{B}\|\|\mathbf{A}\|\|\mathbf{B}\|}{\lambda_{\min}^3(\mathbf{B})} \leq \frac{28\|\mathbf{B}\|^2}{\lambda_{\min}^3(\mathbf{B})} \|\mathbf{A}\|,$$

so we conclude (34) from mean-value theorem.

Moreover, for an arbitrary unit vector \mathbf{v} ,

$$\begin{aligned} & \|\mathcal{H}(\mathbf{x})\mathbf{v} - \mathcal{H}(\mathbf{y})\mathbf{v}\| \\ & \leq 2 \left\| \frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B} + 2(\mathbf{A} \mathbf{x} \mathbf{x}^\top \mathbf{B} - \mathbf{B} \mathbf{x} \mathbf{x}^\top \mathbf{A})}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} \mathbf{v} - \frac{(y^\top \mathbf{B} y)\mathbf{A} - (y^\top \mathbf{A} y)\mathbf{B} + 2(\mathbf{A} y y^\top \mathbf{B} - \mathbf{B} y y^\top \mathbf{A})}{(y^\top \mathbf{B} y)^2} \mathbf{v} \right\| \\ & \quad + 8 \left\| \frac{[(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B}] \mathbf{x} \mathbf{x}^\top \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^3} \mathbf{v} - \frac{[(y^\top \mathbf{B} y)\mathbf{A} - (y^\top \mathbf{A} y)\mathbf{B}] y y^\top \mathbf{B}}{(y^\top \mathbf{B} y)^3} \mathbf{v} \right\| \equiv \text{I} + \text{II}. \end{aligned}$$

Note

$$\begin{aligned} & \nabla \left[\frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B} + 2(\mathbf{A} \mathbf{x} \mathbf{x}^\top \mathbf{B} - \mathbf{B} \mathbf{x} \mathbf{x}^\top \mathbf{A})}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} \mathbf{v} \right] \\ & = \frac{2(\mathbf{x}^\top \mathbf{B})\mathbf{A}\mathbf{v} - 2(\mathbf{x}^\top \mathbf{A})\mathbf{B}\mathbf{v} + 2((\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{v}\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{v})\mathbf{B}) + 2(\mathbf{A} \mathbf{x} \mathbf{v}^\top \mathbf{B} - \mathbf{B} \mathbf{x} \mathbf{v}^\top \mathbf{A})}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} \\ & \quad - 2 \frac{[(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B} + 2(\mathbf{A} \mathbf{x} \mathbf{x}^\top \mathbf{B} - \mathbf{B} \mathbf{x} \mathbf{x}^\top \mathbf{A})] \mathbf{x} \mathbf{v}^\top (2\mathbf{B})}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^3}, \end{aligned}$$

whose norm is bounded by $36\|\mathbf{A}\|\|\mathbf{B}\|^2/\lambda_{\min}^3(\mathbf{B})$, and

$$\begin{aligned} & \nabla \left[\frac{[(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B}] \mathbf{x} \mathbf{x}^\top \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^3} \mathbf{v} \right] = \nabla \left[\frac{(\mathbf{x}^\top \mathbf{B} \mathbf{v}) [(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B}]}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^3} \right] \\ & = \frac{[(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B}] \mathbf{x} \mathbf{v}^\top \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^3} + \frac{(\mathbf{x}^\top \mathbf{B} \mathbf{v}) [(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B} + 2(\mathbf{A} \mathbf{x} \mathbf{x}^\top \mathbf{B} - \mathbf{B} \mathbf{x} \mathbf{x}^\top \mathbf{A})]}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^3} \\ & \quad - 3 \left[\frac{(\mathbf{x}^\top \mathbf{B} \mathbf{v}) [(\mathbf{x}^\top \mathbf{B} \mathbf{x})\mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x})\mathbf{B}] \mathbf{x} \mathbf{x}^\top (2\mathbf{B})}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^4} \right], \end{aligned}$$

whose norm is thus bounded by $20\|\mathbf{A}\|\|\mathbf{B}\|^3/\lambda_{\min}^4(\mathbf{B})$. Again by mean value theorem we have

$$\|\text{I}\| \leq \frac{36\|\mathbf{A}\|\|\mathbf{B}\|^2}{\lambda_{\min}^3(\mathbf{B})} \|\mathbf{x} - \mathbf{y}\|,$$

and

$$\|\Pi\| \leq \frac{20\|\mathbf{A}\|\|\mathbf{B}\|^3}{\lambda_{\min}^4(\mathbf{B})}\|\mathbf{x} - \mathbf{y}\|,$$

so

$$\|\mathcal{H}(\mathbf{x})\mathbf{v} - \mathcal{H}(\mathbf{y})\mathbf{v}\| \leq \frac{56\|\mathbf{A}\|\|\mathbf{B}\|^3}{\lambda_{\min}^4(\mathbf{B})}\|\mathbf{x} - \mathbf{y}\|,$$

which concludes (35) via the definition of operator norm.

Lastly to conclude (36), we utilize the properties of projection matrices, $\|P_{\mathcal{T}(\mathbf{z})}\| \leq 1$, $\|P_{\mathcal{T}(\mathbf{z}_1)} - P_{\mathcal{T}(\mathbf{z}_2)}\| \leq \|\mathbf{z}_1 - \mathbf{z}_2\|$ and hence from matrix operator theory

$$\begin{aligned} & \left\| P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z}')}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')} \right\| \\ & \leq \left\| P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')} \right\| \\ & + \left\| P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z}')} - P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')} \right\| + \left\| P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')} - P_{\mathcal{T}(\mathbf{z}')}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')} \right\| \\ & \leq \|P_{\mathcal{T}(\mathbf{z})}^\top\| \|\mathcal{H}(\mathbf{z})\| \|P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z}')} \| \\ & \quad + \|P_{\mathcal{T}(\mathbf{z})}^\top\| \|\mathcal{H}(\mathbf{z}) - \mathcal{H}(\mathbf{z}')\| \|P_{\mathcal{T}(\mathbf{z}')} \| + \|(P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z}')})^\top\| \|\mathcal{H}(\mathbf{z}')\| \|P_{\mathcal{T}(\mathbf{z}')} \| \\ & \leq L_G \|\mathbf{z} - \mathbf{z}'\| + L_H \|\mathbf{z} - \mathbf{z}'\| + L_G \|\mathbf{z} - \mathbf{z}'\| \\ & = (2L_G + L_H) \|\mathbf{z} - \mathbf{z}'\|. \end{aligned}$$

We complete our proof. ■

We now come to explore what the small gradient condition $\|g(\mathbf{x})\| \leq \gamma$, where $g(\cdot)$ is defined in (32), means for a point \mathbf{x} in the GEV Problem. We first analyze the case where \mathbf{B} is the identity matrix, which reduces to the classical Eigenvector Problem. Define for convenience

$$\gamma_1 \equiv \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \frac{\gamma}{2\lambda_{\text{gap}}}. \quad (37)$$

Lemma 23 When $\mathbf{B} = \mathbf{I}$, we have under $\|\mathbf{w}\| = 1$, an arbitrary constant $\gamma_1 \in (0, 1/2)$ and

$$\left\| \Lambda \mathbf{w} - \frac{\mathbf{w}^\top \Lambda \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \mathbf{w} \right\| \leq \lambda_{\text{gap}} \gamma_1,$$

and for some $j = 1, \dots, d$ (for consistency we define $\lambda_0 = \lambda_1$ and $\lambda_{d+1} = \lambda_d$)

$$\frac{\mathbf{w}^\top \Lambda \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \in \left[\frac{\lambda_{j-1} + \lambda_j}{2}, \frac{\lambda_j + \lambda_{j+1}}{2} \right],$$

together imply

$$(\mathbf{e}_j^\top \mathbf{w})^2 \geq 1 - 4\gamma_1^2.$$

Proof Denote till the rest of this proof $w_i = \mathbf{e}_i^\top \mathbf{w}$. Note we have by

$$\begin{aligned} \lambda_{\text{gap}}^2 \gamma_1^2 & \geq \left\| \Lambda \mathbf{w} - \frac{\mathbf{w}^\top \Lambda \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \mathbf{w} \right\|^2 = \sum_{i=1}^d \left(\lambda_i - \frac{\mathbf{w}^\top \Lambda \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \right)^2 w_i^2 \\ & \geq \sum_{i=1}^{j-1} \left(\lambda_i - \frac{\mathbf{w}^\top \Lambda \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \right)^2 w_i^2 + \sum_{i=j+1}^d \left(\lambda_i - \frac{\mathbf{w}^\top \Lambda \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \right)^2 w_i^2 \\ & \geq \sum_{i=1}^{j-1} \left(\lambda_i - \frac{\lambda_{j-1} + \lambda_j}{2} \right)^2 w_i^2 + \sum_{i=j+1}^d \left(\lambda_i - \frac{\lambda_j + \lambda_{j+1}}{2} \right)^2 w_i^2 \\ & \geq \left(\frac{\lambda_j - \lambda_{j-1}}{2} \right)^2 \sum_{i=1}^{j-1} w_i^2 + \left(\frac{\lambda_{j+1} - \lambda_j}{2} \right)^2 \sum_{i=j+1}^d w_i^2 \geq \frac{\lambda_{\text{gap}}^2}{4} (1 - w_j^2). \end{aligned}$$

This implies the lemma immediately. ■

To study the case of general \mathbf{B} , we first introduce an auxiliary lemma.

Lemma 24 *Given two norms $\|\cdot\|_1, \|\cdot\|_2$ that are equivalent: there are constants $C_L, C_U > 0$ such that for every nonzero vector \mathbf{v} , $C_L\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1 \leq C_U\|\mathbf{v}\|_2$. Then for two given nonzero vectors $\mathbf{w}_1, \mathbf{w}_2$, we have*

$$\left\| \frac{\mathbf{w}_1}{\|\mathbf{w}_1\|_1} - \frac{\mathbf{w}_2}{\|\mathbf{w}_2\|_1} \right\|_1 \leq 2C_L^{-1}C_U \left\| \frac{\mathbf{w}_1}{\|\mathbf{w}_1\|_2} - \frac{\mathbf{w}_2}{\|\mathbf{w}_2\|_2} \right\|_2$$

Proof Without loss of generality we set $\|\mathbf{w}_1\|_2 = 1 = \|\mathbf{w}_2\|_2$. Then using triangle inequality we have

$$\begin{aligned} LHS &= \left\| \frac{\mathbf{w}_1}{\|\mathbf{w}_1\|_1} - \frac{\mathbf{w}_2}{\|\mathbf{w}_2\|_1} \right\|_1 = \frac{\| \|\mathbf{w}_2\|_1 \mathbf{w}_1 - \|\mathbf{w}_1\|_1 \mathbf{w}_2 \|_1}{\|\mathbf{w}_1\|_1 \|\mathbf{w}_2\|_1} \\ &= \frac{\| \|\mathbf{w}_2\|_1 \mathbf{w}_1 - \|\mathbf{w}_1\|_1 \mathbf{w}_1 + \|\mathbf{w}_1\|_1 \mathbf{w}_1 - \|\mathbf{w}_1\|_1 \mathbf{w}_2 \|_1}{\|\mathbf{w}_1\|_1 \|\mathbf{w}_2\|_1} \\ &\leq \frac{\| \|\mathbf{w}_2\|_1 - \|\mathbf{w}_1\|_1 \| \|\mathbf{w}_1\|_1 + \|\mathbf{w}_1\|_1 \|\mathbf{w}_1 - \mathbf{w}_2\|_1}{\|\mathbf{w}_1\|_1 \|\mathbf{w}_2\|_1} \\ &\leq \frac{2\|\mathbf{w}_1 - \mathbf{w}_2\|_1 \|\mathbf{w}_1\|_1}{\|\mathbf{w}_1\|_1 \|\mathbf{w}_2\|_1} = 2\|\mathbf{w}_2\|_1^{-1} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|_1 \leq 2C_L^{-1}\|\mathbf{w}_2\|_2^{-1} \cdot C_U\|\mathbf{w}_1 - \mathbf{w}_2\|_2 = RHS. \end{aligned}$$

We conclude the following lemma. ■

Lemma 25 *We have for $\mathbf{x} \in \mathcal{S}^{d-1}$,*

$$\gamma \in \left(0, \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{-1/2} \lambda_{\text{gap}} \right), \quad (38)$$

and $\|g(\mathbf{x})\| \leq \gamma$ implies that there exists at least one $j = 1, \dots, d$ such that

$$(\mathbf{e}_j^\top \mathbf{B}^{1/2} \mathbf{x})^2 \geq (1 - 4\gamma_1^2) \|\mathbf{B}^{1/2} \mathbf{x}\|^2. \quad (39)$$

Furthermore, we have that there exists at least one $j = 1, \dots, d$ such that

$$\min(\|\mathbf{x} - \mathbf{v}_j\|, \|\mathbf{x} + \mathbf{v}_j\|) \leq 4\sqrt{2} \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \cdot \gamma_1. \quad (40)$$

Proof Since \mathbf{B} is positive definite, letting in (23) $\mathbf{w} = \mathbf{B}^{1/2} \mathbf{x} / \|\mathbf{B}^{1/2} \mathbf{x}\|$, we have $\|\mathbf{w}\| = 1$ and recall that $\mathbf{A} = \mathbf{B}^{1/2} \mathbf{\Lambda} \mathbf{B}^{1/2}$

$$\begin{aligned} \left\| \mathbf{\Lambda} \mathbf{w} - \frac{\mathbf{w}^\top \mathbf{\Lambda} \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \mathbf{w} \right\| &= \|\mathbf{B}^{1/2} \mathbf{x}\| \left\| \mathbf{B}^{-1/2} \left(\frac{\mathbf{B}^{1/2} \mathbf{\Lambda} \mathbf{B}^{1/2}}{\mathbf{x}^\top \mathbf{B} \mathbf{x}} \mathbf{x} - \frac{\mathbf{x}^\top \mathbf{B}^{1/2} \mathbf{\Lambda} \mathbf{B}^{1/2} \mathbf{x}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} \mathbf{B} \mathbf{x} \right) \right\| \\ &\leq \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \left\| \frac{(\mathbf{x}^\top \mathbf{B} \mathbf{x}) \mathbf{A} - (\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{B}}{(\mathbf{x}^\top \mathbf{B} \mathbf{x})^2} \mathbf{x} \right\| \leq \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \frac{\gamma}{2} = \lambda_{\text{gap}} \gamma_1. \end{aligned}$$

Note (41) gives $\gamma_1 \in (0, 1/2)$, and hence applying Lemma 23 gives the following: there is at least one $j = 1, \dots, d$ such that $(\mathbf{e}_j^\top \mathbf{w})^2 \geq 1 - 4\gamma_1^2$. Translating this back in terms of \mathbf{x} concludes (39).

To conclude (42) we note (39) gives if $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle_{\mathbf{B}} \equiv \mathbf{z}_1^\top \mathbf{B} \mathbf{z}_2$ and $\|\mathbf{z}\|_{\mathbf{B}} \equiv \langle \mathbf{z}, \mathbf{z} \rangle_{\mathbf{B}}^{1/2}$:

$$\left\langle \mathbf{e}_j^\top \mathbf{B}^{-1/2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}} \right\rangle_{\mathbf{B}}^2 \geq 1 - 4\gamma_1^2,$$

so

$$\begin{aligned} \left\| \mathbf{e}_j^\top \mathbf{B}^{-1/2} \pm \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}} \right\|_{\mathbf{B}}^2 &= \left\| \mathbf{e}_j^\top \mathbf{B}^{-1/2} \right\|_{\mathbf{B}}^2 + \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}} \right\|_{\mathbf{B}}^2 \pm 2 \left\langle \mathbf{e}_j^\top \mathbf{B}^{-1/2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}} \right\rangle_{\mathbf{B}} \\ &= 2 \pm 2 \left\langle \mathbf{e}_j^\top \mathbf{B}^{-1/2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}} \right\rangle_{\mathbf{B}}, \end{aligned}$$

and hence using $1 - \sqrt{1-t} \leq t$ for $t \in [0, 1]$

$$\min \left\| \mathbf{e}_j^\top \mathbf{B}^{-1/2} \pm \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}} \right\|_{\mathbf{B}}^2 = 2 - 2 \left| \left\langle \mathbf{e}_j^\top \mathbf{B}^{-1/2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}} \right\rangle_{\mathbf{B}} \right| = 2 - 2\sqrt{1 - 4\gamma_1^2} \leq 8\gamma_1^2.$$

Using this and applying Lemma 24 with $\|\cdot\|_1 = \|\cdot\|$ and $\|\cdot\|_2 = \|\cdot\|_{\mathbf{B}}$ we have $\lambda_{\max}^{-1/2}(\mathbf{B})\|\mathbf{v}\|_{\mathbf{B}} \leq \|\mathbf{v}\| \leq \lambda_{\min}^{-1/2}(\mathbf{B})\|\mathbf{v}\|_{\mathbf{B}}$ and hence for two given nonzero vectors (in the Euclidean norm) \mathbf{x} and $\mp \mathbf{v}_j = \mp \|\mathbf{e}_j^\top \mathbf{B}^{-1/2}\|^{-1} \mathbf{e}_j^\top \mathbf{B}^{-1/2}$

$$\min \|\mathbf{x} \pm \mathbf{v}_j\| \leq 2 \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \cdot \min \left\| \mathbf{e}_j^\top \mathbf{B}^{-1/2} \pm \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}} \right\|_{\mathbf{B}} \leq 4\sqrt{2} \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \cdot \gamma_1.$$

■

Now we finish the proof of Lemma 3.

Proof [Proof of Lemma 3]

- (i) We have $\mathbf{A}\mathbf{x} = (\mathbf{x}^\top \mathbf{A}\mathbf{x} / \mathbf{x}^\top \mathbf{B}\mathbf{x})\mathbf{B}\mathbf{x}$ if and only if $g(\mathbf{x}) = 0$. For $\Lambda = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ being WLOG diagonal, one can see that for $j = 2, \dots, d$ and \mathbf{v}_j on the unit sphere with $\mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{B}\mathbf{v}_j$,

$$\mathcal{H}(\mathbf{v}_j) = -2 \cdot \frac{(\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j)\mathbf{A} - (\mathbf{v}_j^\top \mathbf{A}\mathbf{v}_j)\mathbf{B}}{(\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j)^2} = -2 \cdot \frac{\mathbf{A} - \lambda_j \mathbf{B}}{\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j}.$$

Thus

$$\begin{aligned} & (\mathbf{v}_1 - c\mathbf{v}_j)^\top \mathcal{H}(\mathbf{v}_j)(\mathbf{v}_1 - c\mathbf{v}_j) \\ &= -2 \cdot \frac{(\mathbf{v}_1 - c\mathbf{v}_j)^\top (\mathbf{A} - \lambda_j \mathbf{B})(\mathbf{v}_1 - c\mathbf{v}_j)}{\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j} \\ &= -2 \cdot \frac{\mathbf{v}_1^\top (\mathbf{A} - \lambda_j \mathbf{B})\mathbf{v}_1}{\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j} + 4c \cdot \frac{\mathbf{v}_j^\top (\mathbf{A} - \lambda_j \mathbf{B})\mathbf{v}_1}{\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j} - 2c^2 \cdot \frac{\mathbf{v}_j^\top (\mathbf{A} - \lambda_j \mathbf{B})\mathbf{v}_j}{\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j} \\ &= -2 \cdot \frac{\mathbf{v}_1^\top (\mathbf{A} - \lambda_j \mathbf{B})\mathbf{v}_1}{\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j} \\ &= -2(\lambda_1 - \lambda_j) \cdot \frac{\mathbf{v}_1^\top \mathbf{B}\mathbf{v}_1}{\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j} \leq -2(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|}. \end{aligned}$$

In the display above, we use the fact that $\mathbf{v}_j^\top (\mathbf{A} - \lambda_j \mathbf{B})\mathbf{v}_1 = (\lambda_1 - \lambda_j)\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_1 = 0$ and $\mathbf{v}_j^\top (\mathbf{A} - \lambda_j \mathbf{B})\mathbf{v}_j = (\lambda_j - \lambda_j)\mathbf{v}_j^\top \mathbf{B}\mathbf{v}_j = 0$. By picking $c = \mathbf{v}_j^\top \mathbf{v}_1$ such that $P_{\mathcal{T}(\mathbf{v}_j)}\mathbf{v}_1 = \mathbf{v}_1 - (\mathbf{v}_j^\top \mathbf{v}_1)\mathbf{v}_j = \mathbf{v}_1 - c\mathbf{v}_j$, we conclude

$$\|P_{\mathcal{T}(\mathbf{v}_j)}\mathbf{v}_1\| = \sqrt{1 + (\mathbf{v}_j^\top \mathbf{v}_1)^2 - 2(\mathbf{v}_j^\top \mathbf{v}_1)^2} = \sqrt{1 - (\mathbf{v}_j^\top \mathbf{v}_1)^2} \in (0, 1],$$

(since $\mathbf{v}_1 \neq \pm \mathbf{v}_j$ otherwise $0 = \mathbf{v}_j^\top \mathbf{B}\mathbf{v}_1 = \pm \mathbf{v}_1^\top \mathbf{B}\mathbf{v}_1$ which leads to $\mathbf{v}_1 = 0$ due to the positive definiteness of \mathbf{B} .) and hence from the above two displays

$$\mathbf{v}_1^\top \left[P_{\mathcal{T}(\mathbf{v}_j)}^\top \mathcal{H}(\mathbf{v}_j) P_{\mathcal{T}(\mathbf{v}_j)} \right] \mathbf{v}_1 \leq -2(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|} \|P_{\mathcal{T}(\mathbf{v}_j)}(\mathbf{v}_1)\|^2.$$

- (ii) To conclude points that are close to $P_{\mathcal{T}(\mathbf{v}_j)}\mathbf{v}_1$, Lemma 25 gives for $\mathbf{x} \in \mathcal{S}^{d-1}$,

$$\gamma \in \left(0, \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{-1/2} \lambda_{\text{gap}} \right), \quad (41)$$

and $\|g(\mathbf{x})\| \leq \gamma$ implies that there exists at least one $j = 1, \dots, d$ such that

$$\min \|\mathbf{x} \pm \mathbf{v}_j\| \leq 4\sqrt{2} \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \cdot \gamma_1. \quad (42)$$

Without loss of generality we suppose the minus sign in the above display is taken, so $\min \|\mathbf{x} - \mathbf{v}_j\| \leq 4\sqrt{2} \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \cdot \gamma_1$. Then given the definition of γ_1 in (37) we have from Lemma 22 that

$$\begin{aligned}
& \mathbf{v}_1^\top \left[P_{\mathcal{T}(\mathbf{x})}^\top \mathcal{H}(\mathbf{x}) P_{\mathcal{T}(\mathbf{x})} \right] \mathbf{v}_1 \\
& \leq \mathbf{v}_1^\top \left[P_{\mathcal{T}(\mathbf{v}_j)}^\top \mathcal{H}(\mathbf{v}_j) P_{\mathcal{T}(\mathbf{v}_j)} \right] \mathbf{v}_1 + \left\| P_{\mathcal{T}(\mathbf{v}_j)}^\top \mathcal{H}(\mathbf{v}_j) P_{\mathcal{T}(\mathbf{v}_j)} - P_{\mathcal{T}(\mathbf{x})}^\top \mathcal{H}(\mathbf{x}) P_{\mathcal{T}(\mathbf{x})} \right\| \\
& \leq -2(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|} + (2L_G + L_H) \|\mathbf{x} - \mathbf{v}_j\| \\
& \leq -(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|} \leq -(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|} \|P_{\mathcal{T}(\mathbf{x})} \mathbf{v}_1\|^2,
\end{aligned}$$

as long as (combined with (42))

$$4\sqrt{2}(2L_G + L_H) \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})} \right)^{1/2} \cdot \gamma_1 \leq (\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|},$$

where we applied $\|P_{\mathcal{T}(\mathbf{x})} \mathbf{v}_1\| \leq 1$. This completes the proof of Lemma combining with the definition of β in (4). ■

E DEFERRED PROOFS OF §B.1

We collect the deferred proofs from §B.1.

E.1 PROOF OF LEMMA 4

Proof [Proof of Lemma 4] Since $M = \mathcal{V} \log^{\frac{1}{\alpha}} \epsilon^{-1}$, we have from Assumption ?? that for each $t \geq 1$,

$$\begin{aligned}
\mathbb{P}(\|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\| > M) &= \mathbb{P}\left(\exp\left(\frac{\|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\|^\alpha}{\mathcal{V}^\alpha}\right) > \exp\left(\frac{M^\alpha}{\mathcal{V}^\alpha}\right)\right) \\
&\leq \exp\left(-\frac{M^\alpha}{\mathcal{V}^\alpha}\right) \mathbb{E} \exp\left(\frac{\|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\|^\alpha}{\mathcal{V}^\alpha}\right) \leq 2\epsilon.
\end{aligned}$$

where we apply the Markov inequality and Assumption ?? (with law of total expectation applied). Taking a union bound,

$$\mathbb{P}(\mathcal{T}_M \leq T_\eta^*) \leq \sum_{t=1}^{T_\eta^*} \mathbb{P}(\|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\| > M) \leq 2T_\eta^* \epsilon.$$
■

E.2 PROOF OF LEMMA 5

Proof [Proof of Lemma 5] For all $\mathbf{u}, \mathbf{v} \in \mathcal{S}^{d-1}$, we have

$$\begin{aligned}
\|g(\mathbf{u}) - g(\mathbf{v})\| &\leq \|\mathbf{I} - \mathbf{u}\mathbf{u}^\top\| \|\nabla F(\mathbf{u}) - \nabla F(\mathbf{v})\| + \|\mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top\| \|\nabla F(\mathbf{v})\| \\
&\leq 1 \cdot L_K \|\mathbf{u} - \mathbf{v}\| + 2\|\mathbf{u} - \mathbf{v}\| \cdot L_F \\
&= (L_K + 2L_F) \|\mathbf{u} - \mathbf{v}\|, \\
\|\mathcal{H}(\mathbf{u}) - \mathcal{H}(\mathbf{v})\| &\leq \|\nabla^2 F(\mathbf{u}) - \nabla^2 F(\mathbf{v})\| + (\|\mathbf{u} - \mathbf{v}\| \|\nabla F(\mathbf{u})\| + \|\mathbf{v}\| \|\nabla F(\mathbf{u}) - \nabla F(\mathbf{v})\|) \|\mathbf{I}\| \\
&\leq L_Q \|\mathbf{u} - \mathbf{v}\| + (\|\mathbf{u} - \mathbf{v}\| \cdot L_F + 1 \cdot L_K \|\mathbf{u} - \mathbf{v}\|) \cdot 1 \\
&= (L_Q + L_F + L_K) \|\mathbf{u} - \mathbf{v}\|, \\
\|\mathcal{N}(\mathbf{u}) - \mathcal{N}(\mathbf{v})\| &\leq \|\mathbf{u} - \mathbf{v}\| (\|\nabla F(\mathbf{u})\| + \|\nabla^2 F(\mathbf{u})\| \|\mathbf{u}\|) \\
&\quad + \|\mathbf{v}\| (\|\nabla F(\mathbf{u}) - \nabla F(\mathbf{v})\| + \|\nabla^2 F(\mathbf{u}) - \nabla^2 F(\mathbf{v})\| \|\mathbf{u}\| + \|\nabla^2 F(\mathbf{v})\| \|\mathbf{u} - \mathbf{v}\|) \\
&\leq \|\mathbf{u} - \mathbf{v}\| (L_F + L_K \cdot 1) + 1 \cdot (L_K \|\mathbf{u} - \mathbf{v}\| + L_Q \|\mathbf{u} - \mathbf{v}\| \cdot 1 + L_K \cdot \|\mathbf{u} - \mathbf{v}\|) \\
&= (L_F + 3L_K + L_Q) \|\mathbf{u} - \mathbf{v}\|, \\
\|\mathcal{H}(\mathbf{v})\| &\leq \|\nabla^2 F(\mathbf{v})\| + \|\mathbf{v}\| \|\nabla F(\mathbf{v})\| \|\mathbf{I}\| \leq L_K + 1 \cdot L_F \cdot 1 = L_K + L_F.
\end{aligned}$$

which implies that $g(\mathbf{v})$ is $(L_G \equiv L_K + 2L_F)$ -Lipschitz, $\mathcal{H}(\mathbf{v})$ is $(L_H \equiv L_Q + L_F + L_K)$ -Lipschitz, $\mathcal{N}(\mathbf{v})$ is $(L_N \equiv L_F + 3L_K + L_Q)$ -Lipschitz and $\|\mathcal{H}(\mathbf{v})\| \leq B_H \equiv L_F + L_K$ within $\{\mathbf{v} : \|\mathbf{v}\| \leq 1, \|\mathbf{v} - \mathbf{v}^*\| \leq \delta\}$. \blacksquare

E.3 PROOF OF LEMMA 6

Proof [Proof of Lemma 6] We have by a Taylor series expansion that for any $y \in \mathbb{R}$ satisfying $|y| \leq 1/2$

$$\left| (1 - y)^{-1/2} - 1 - \frac{y}{2} \right| \leq \frac{3y^2}{8} \sum_{k=0}^{\infty} |y|^k \leq \frac{3y^2}{4}.$$

When $\eta \leq 1/(5M)$, on the event $(\|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\| \leq M)$, by letting $y = 2\eta \mathbf{v}_{t-1}^\top \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t) - \eta^2 \|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\|^2$ we have

$$|y| \leq 2\eta \|\mathbf{v}_{t-1}^\top \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\| + \eta^2 \|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\|^2 \leq 2\eta M + \eta^2 M^2 \leq (11/5)\eta M < 1/2,$$

and hence combining the above two displays gives

$$\begin{aligned}
&\left| \|\mathbf{v}_{t-1} - \eta \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\|^{-1} - 1 - \eta \mathbf{v}_{t-1}^\top \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t) \right| \\
&\leq \left| (1 - 2\eta \mathbf{v}_{t-1}^\top \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t) + \eta^2 \|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\|^2)^{-1/2} - 1 - \eta \mathbf{v}_{t-1}^\top \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t) \right| \\
&\leq \left| (1 - y)^{-1/2} - 1 - \frac{y}{2} \right| + \frac{\eta^2 \|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\|^2}{2} \\
&\leq \frac{3y^2}{4} + \frac{1}{2} \eta^2 M^2 \leq \frac{3}{4} \cdot \frac{121}{25} \eta^2 M^2 + \frac{1}{2} \eta^2 M^2 \leq 5\eta^2 M^2.
\end{aligned} \tag{43}$$

By defining

$$\boldsymbol{\xi}_t = (\mathbf{I} - \mathbf{v}_{t-1} \mathbf{v}_{t-1}^\top) (\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t) - D(\mathbf{v}_{t-1}) \nabla F(\mathbf{v}_{t-1})), \tag{44}$$

and

$$\begin{aligned}
\mathbf{Q}_t &= \eta^{-2} \cdot (\|\mathbf{v}_{t-1} - \eta \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)\|^{-1} - 1 - \eta \mathbf{v}_{t-1}^\top \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)) (\mathbf{v}_{t-1} - \eta \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)) \\
&\quad - (\mathbf{v}_{t-1}^\top \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t)) \Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t),
\end{aligned} \tag{45}$$

the update formula (??) is equivalent to

$$\mathbf{v}_t = \mathbf{v}_{t-1} - \eta D(\mathbf{v}_{t-1}) g(\mathbf{v}_{t-1}) + \eta \boldsymbol{\xi}_t + \eta^2 \mathbf{Q}_t. \tag{46}$$

Using (43), we have

$$\|\mathbf{Q}_t\| \leq \eta^{-2} \cdot 5\eta^2 M^2 \cdot (1 + \eta M) + M^2 \leq 7M^2.$$

Recall that we denote $D = D(\mathbf{v}^*)$, $\mathcal{H}_* = \mathcal{H}(\mathbf{v}^*)$, $\mathcal{N}_* = \mathcal{N}(\mathbf{v}^*)$. By defining

$$\mathbf{R}_t = D(\mathcal{H}_* + \mathcal{N}_*)(\mathbf{v}_{t-1} - \mathbf{v}^*) - D(\mathbf{v}_{t-1})g(\mathbf{v}_{t-1}), \quad (47)$$

we have

$$\mathbf{v}_t = \mathbf{v}_{t-1} - \eta D(\mathcal{H}_* + \mathcal{N}_*)(\mathbf{v}_{t-1} - \mathbf{v}^*) + \eta \boldsymbol{\xi}_t + \eta \mathbf{R}_t + \eta^2 \mathbf{Q}_t.$$

Since $(\mathbf{I} - \mathbf{v}_{t-1} \mathbf{v}_{t-1}^\top)$ is \mathcal{F}_{t-1} -measurable, we know that $\mathbb{E}[\boldsymbol{\xi}_t | \mathcal{F}_{t-1}] = 0$ and hence $\{\boldsymbol{\xi}_t\}$ is a vector-valued martingale difference sequence. Additionally, we have $\|\mathbf{I} - \mathbf{v}_{t-1} \mathbf{v}_{t-1}^\top\| \leq 1$, and hence from Assumption ?? and Lemma 18 we know

$$\mathbb{E} \exp \left(\frac{\|\boldsymbol{\xi}_t\|^\alpha}{(G_\alpha \mathcal{V})^\alpha} \right) \leq \mathbb{E} \exp \left(\frac{\|\Gamma(\mathbf{v}_{t-1}; \boldsymbol{\zeta}_t) - D(\mathbf{v}_{t-1}) \nabla F(\mathbf{v}_{t-1})\|^\alpha}{(G_\alpha \mathcal{V})^\alpha} \right) \leq 2$$

which implies that $\boldsymbol{\xi}$ is α -sub-Weibull with parameter $G_\alpha \mathcal{V}$.

Finally, we apply the mean-value theorem using (10) and $g(\mathbf{v}^*) = 0$ to obtain

$$\begin{aligned} \|\mathbf{R}_t\| &= \|D(\mathcal{H}_* + \mathcal{N}_*)(\mathbf{v}_{t-1} - \mathbf{v}^*) - D(\mathbf{v}_{t-1})g(\mathbf{v}_{t-1})\| \\ &\leq D \left\| (\mathcal{H}_* + \mathcal{N}_*)(\mathbf{v}_{t-1} - \mathbf{v}^*) - \int_0^1 \mathcal{H}(\mathbf{v}^* + \theta(\mathbf{v}_{t-1} - \mathbf{v}^*)) + \mathcal{N}(\mathbf{v}^* + \theta(\mathbf{v}_{t-1} - \mathbf{v}^*)) d\theta (\mathbf{v}_{t-1} - \mathbf{v}^*) \right\| \\ &\quad + \|D - D(\mathbf{v}_{t-1})\| \|g(\mathbf{v}_{t-1})\| \\ &\leq D(L_H + L_N) \|\mathbf{v}_{t-1} - \mathbf{v}^*\|^2 + L_D L_G \|\mathbf{v}_{t-1} - \mathbf{v}^*\|^2 \end{aligned}$$

where we use the Lipschitz continuity of $D(\mathbf{v})$, $g(\mathbf{v})$, $\mathcal{H}(\mathbf{v})$, $\mathcal{N}(\mathbf{v})$. This completes the proof of Lemma 6. \blacksquare

E.4 PROOF OF LEMMA 7

Proof [Proof of Lemma 7] Under initialization condition (??), we have the following:

(i) For all unit vector \mathbf{v} , since $\|\mathbf{v}\| = \|\mathbf{v}^*\| = 1$ we have

$$\|(\mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v} - \mathbf{v}^*)\| = -\mathbf{v}^{*\top}(\mathbf{v} - \mathbf{v}^*) = \frac{1}{2} \|\mathbf{v}\|^2 - \mathbf{v}^{*\top} \mathbf{v} + \frac{1}{2} \|\mathbf{v}^*\|^2 = \frac{1}{2} \|\mathbf{v} - \mathbf{v}^*\|^2.$$

Because

$$\left((\mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v} - \mathbf{v}^*) \right)^\top \left((\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v} - \mathbf{v}^*) \right) = 0,$$

by the Pythagorean theorem we have

$$\|(\mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v} - \mathbf{v}^*)\|^2 + \|(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v} - \mathbf{v}^*)\|^2 = \|\mathbf{v} - \mathbf{v}^*\|^2$$

Combining the above equalities and plugging in $\mathbf{v} = \mathbf{v}_t$ gives

$$\|\boldsymbol{\Delta}_t\|^2 = \|\mathbf{v}_t - \mathbf{v}^*\|^2 - \frac{1}{4} \|\mathbf{v}_t - \mathbf{v}^*\|^4,$$

which admits the following solution given $\mathbf{v}_t^\top \mathbf{v}^* \geq 0$:

$$\|\mathbf{v}_t - \mathbf{v}^*\|^2 = 2 - \sqrt{4 - 4\|\boldsymbol{\Delta}_t\|^2},$$

and hence

$$\|\boldsymbol{\Delta}_t\|^2 \leq \|\mathbf{v}_t - \mathbf{v}^*\|^2 = \frac{4\|\boldsymbol{\Delta}_t\|^2}{2 + \sqrt{4 - 4\|\boldsymbol{\Delta}_t\|^2}} \leq 2\|\boldsymbol{\Delta}_t\|^2.$$

(ii) Under initialization condition (??), for all $\mathbf{u} \in \mathcal{T}(\mathbf{v}^*)$, we have $\mathbf{u}^\top \mathcal{H}_* \mathbf{u} \geq \mu \|\mathbf{u}\|^2$. Hence for $\eta \leq 1/(DB_H)$, we have

$$\|(\mathbf{I} - \eta D\mathcal{M}_*)^{1/2} \mathbf{u}\| \leq (1 - \eta D\mu)^{1/2} \|\mathbf{u}\|. \quad (48)$$

By noticing that $(\mathbf{I} - \eta D\mathcal{M}_*)^{(t-1)/2} \mathbf{u} \in \mathcal{T}(\mathbf{v}^*)$, for all $t \geq 1$, we could inductively plug in $(\mathbf{I} - \eta D\mathcal{M}_*)^{(t-1)/2} \mathbf{u}$ to \mathbf{u} in (48) and obtain for each $t \geq 0$

$$\|(\mathbf{I} - \eta D\mathcal{M}_*)^t \mathbf{u}\| \leq (1 - \eta D\mu)^t \|\mathbf{u}\|.$$

\blacksquare

E.5 PROOF OF LEMMA 8

Proof [Proof of Lemma 8] By left multiplying (12) in Lemma 6 by $(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top})$ and noticing $(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathcal{N}_* = 0$, we obtain

$$\begin{aligned} \Delta_t &= \Delta_{t-1} - \eta D(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathcal{H}_*(\mathbf{v}_{t-1} - \mathbf{v}^*) + \eta(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \xi_t \\ &\quad + \eta(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathbf{R}_t + \eta^2(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathbf{Q}_t. \end{aligned}$$

We have the decomposition

$$(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathcal{H}_*(\mathbf{v}_{t-1} - \mathbf{v}^*) = (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathcal{H}_* \Delta_t + (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathcal{H}_* \cdot (\mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v}_{t-1} - \mathbf{v}^*),$$

where $(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathcal{H}_* \Delta_t = \mathcal{M}_* \Delta_t$, and based on Lemma 7 and $\|\mathcal{H}_*\| \leq B_H$,

$$\|(\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathcal{H}_* \cdot (\mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v}_{t-1} - \mathbf{v}^*)\| \leq \frac{B_H}{2} \|\mathbf{v}_{t-1} - \mathbf{v}^*\|^2.$$

We set

$$\begin{aligned} \chi_t &= (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \xi_t, \\ \mathbf{S}_t &= (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathbf{R}_t - D \cdot (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathcal{H}_* \cdot (\mathbf{v}^* \mathbf{v}^{*\top})(\mathbf{v}_{t-1} - \mathbf{v}^*), \\ \mathbf{P}_t &= (\mathbf{I} - \mathbf{v}^* \mathbf{v}^{*\top}) \mathbf{Q}_t. \end{aligned}$$

Then by combining all of the results above, we have

$$\Delta_t = (\mathbf{I} - \eta D \mathcal{M}_*) \Delta_{t-1} + \eta \chi_t + \eta \mathbf{S}_t + \eta^2 \mathbf{P}_t,$$

which proves (17). The rest of Lemma 8 can be easily verified in steps similar to the proof of Lemma 6. ■

E.6 PROOF OF LEMMA 9

Proof [Proof of Lemma 9] For $t = 0$ the lemma holds by definition. In general if it holds for $t - 1$ then from the definitions in (18) we have on $(t < \mathcal{T}_M)$ that $\tilde{\mathbf{S}}_s = \mathbf{S}_s, \tilde{\mathbf{P}}_s = \mathbf{P}_s$ for all $s \leq t$, so the conclusion holds for t . Iteratively applying (19) we obtain (20), which concludes our lemma. ■

E.7 PROOF OF LEMMA 10

Proof [Proof of Lemma 10] For any fixed $t \geq 0$, we have the following:

(i) For the first term on the right hand of (20) which we repeat here

$$\begin{aligned} \overline{\Delta}_t &= (\mathbf{I} - \eta D \mathcal{M}_*)^t \Delta_0 + \eta \sum_{s=1}^t (\mathbf{I} - \eta D \mathcal{M}_*)^{t-s} \chi_s \\ &\quad + \eta \sum_{s=1}^t (\mathbf{I} - \eta D \mathcal{M}_*)^{t-s} \tilde{\mathbf{S}}_s + \eta^2 \sum_{s=1}^t (\mathbf{I} - \eta D \mathcal{M}_*)^{t-s} \tilde{\mathbf{P}}_s, \end{aligned} \tag{20}$$

since $\chi_s \in \mathcal{T}(\mathbf{v}^*)$, (16) in Lemma 7 implies $\|(\mathbf{I} - \eta D \mathcal{M}_*)^{t-s} \chi_s\| \leq (1 - \eta D \mu)^{t-s} \|\chi_s\|$. Hence we have $\|(\mathbf{I} - \eta D \mathcal{M}_*)^{t-s} \chi_s\|_{\psi_\alpha} \leq (1 - \eta D \mu)^{t-s} \|\chi_s\|_{\psi_\alpha} \leq (1 - \eta D \mu)^{t-s} G_\alpha \mathcal{V}$ and

$$\sum_{s=1}^t \|(\mathbf{I} - \eta D \mathcal{M}_*)^{t-s} \chi_s\|_{\psi_\alpha}^2 \leq \eta^2 \sum_{s=1}^t (1 - \eta D \mu)^{2(t-s)} G_\alpha^2 \mathcal{V}^2 \leq \frac{G_\alpha^2 \mathcal{V}^2}{D \mu} \cdot \eta$$

Modifying the results in Fan et al. [2012] provides a concentration inequality for α -sub-Weibull random vectors, which gives¹

$$\mathbb{P} \left(\left\| \eta \sum_{s=1}^t (\mathbf{I} - \eta D \mathcal{M}_*)^{t-s} \chi_s \right\| \geq \frac{8 G_\alpha \mathcal{V}}{\sqrt{D \mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} \right) \leq \left(12 + 8 \left(\frac{3}{\alpha} \right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) \epsilon.$$

¹A similar concentration inequality method for the scalar case is adopted by Li and Jordan [2021].

(ii) For the second term on the right-hand side of (20), by applying (16) in Lemma 7 and using Lemma 8, given $\|v_{s-1} - v^*\| \leq r$ for all $s = 1, \dots, t$ we have,

$$\left\| \eta \sum_{s=1}^t (\mathbf{I} - \eta D\mathcal{M}_*)^{t-s} \tilde{\mathbf{S}}_s \right\| \leq \eta \sum_{s=1}^t (1 - \eta D\mu)^{t-s} \cdot \rho r^2 \leq \frac{\rho r^2}{D\mu}. \quad (49)$$

(iii) For the third term on the right-hand side of (20), from Lemma 8 we know $\|\tilde{\mathbf{P}}_t\| \leq 7M^2$ and

$$\left\| \eta^2 \sum_{s=1}^t (\mathbf{I} - \eta D\mathcal{M}_*)^{t-s} \tilde{\mathbf{P}}_s \right\| \leq \eta^2 \sum_{s=1}^t (1 - \eta D\mu)^{t-s} \cdot 7M^2 = \frac{7\mathcal{V}^2}{D\mu} \log^{\frac{2}{\alpha}} \epsilon^{-1} \cdot \eta,$$

where we use the definition of M in (8).

The lemma is concluded by combining the above three items and taking union bound on probability. \blacksquare

E.8 PROOF OF LEMMA 11

Proof [Proof of Lemma 11] From the given assumptions, under scaling condition (??), we have

$$r = 2 \max \left\{ \|\Delta_0\|, \frac{2^7 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} \right\} \leq \frac{D\mu}{16\rho}.$$

We let event \mathcal{J} be (21) holding for each $t \in [0, T]$, i.e.

$$\|\bar{\Delta}_t - (\mathbf{I} - \eta D\mathcal{M}_*)^t \Delta_0\| \leq \frac{8G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} + \frac{7\mathcal{V}^2}{D\mu} \log^{\frac{2}{\alpha}} \epsilon^{-1} \cdot \eta.$$

Then on event \mathcal{J} , under scaling condition (??), because $\|\Delta_0\| \leq \frac{r}{2}$, for each $t \in [0, T]$ we have

$$\|\bar{\Delta}_t\| \leq \|\Delta_0\| + \frac{16G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} \leq \frac{r}{2} + \frac{r}{16} + \frac{r}{16} \leq r.$$

Applying Lemma 10 and taking a union bound gives

$$\mathbb{P}(\mathcal{J}) \geq 1 - \left(12 + 8 \left(\frac{3}{\alpha} \right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) T\epsilon.$$

Furthermore, using (16) in Lemma 7 and definition of T_η^* in (??), if $T_\eta^* \in [0, T]$, on event \mathcal{J} we have at time T_η^*

$$\|\bar{\Delta}_{T_\eta^*}\| \leq \|(\mathbf{I} - \eta D\mathcal{M}_*)^{T_\eta^*} \Delta_0\| + \frac{16G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} \leq \frac{r}{8} + \frac{r}{16} + \frac{r}{16} \leq \frac{r}{4}.$$

In Lemma 9 we have shown that, on the event $(T < \mathcal{T}_M)$, we have $\bar{\Delta}_t = \Delta_t$. In Lemma 4, we have proved $\mathbb{P}(T < \mathcal{T}_M) \geq 1 - 2T\epsilon$. Together with Lemma 10, we take an intersection and obtain

$$\mathbb{P}(\mathcal{J} \cap (T < \mathcal{T}_M)) \geq 1 - \mathbb{P}(\mathcal{J}^c) - \mathbb{P}(T \geq \mathcal{T}_M) \geq 1 - \left(14 + 8 \left(\frac{3}{\alpha} \right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) T\epsilon.$$

At this point we have proved all elements in Lemma 11. \blacksquare

References

- Xiequan Fan, Ion Grama, and Quansheng Liu. Large deviation exponential inequalities for supermartingales. *Electronic Communications in Probability*, 17, 2012.
- Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.
- Chris Junchi Li and Michael I Jordan. Stochastic approximation for online tensorial independent component analysis. In *Conference on Learning Theory*, pages 3051–3106. PMLR, 2021.
- Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer, 2006.