Appendix A. Oja’s algorithm and Krasulina’s method

In this section, we briefly discuss two classic first-order algorithms for eigenproblem, Oja’s method and Krasulina’s method. Let $w_t \in \mathbb{R}^d$ be the iterates of estimating the top eigenvector of $A_n$ at time $t$ and $\eta$ denotes the step size. Oja’s algorithm has the following update rule:

$$w_{t+1}' = w_t + \eta A w_t, \quad w_{t+1} = \frac{w_{t+1}'}{\|w_{t+1}'\|}.$$  

Hence, Oja’s method is equivalent to the power method when $\eta \to \infty$. Krasulina’s method uses a similar update rule to Oja’s update but has an additional term:

$$w_{t+1}' = w_t + \eta (I_d - w_t w_t^\top) A w_t, \quad w_{t+1} = \frac{w_{t+1}'}{\|w_{t+1}'\|},$$

which is the gradient descent on the objective function:

$$\min_w \frac{1}{n} \sum_{t=1}^n \|a_t - \frac{w w^\top}{\|w\|^2} a_t \|^2.$$  

The operator $(I_d - w_t w_t^\top)$ is the projection operator, and Amid and Warmuth (2020) showed that Krasulina’s update corresponds to a projected gradient descent step on the Stiefel manifold $\text{St}(d, 1) = \{w \in \mathbb{R}^d : w^\top w = 1\}$. Thus, Krasulina’s method is a Riemannian optimization method, which is reviewed in the following section. Note that there are fundamental differences between Oja’s algorithm and Krasulina’s methods. Krasulina’s method is equivalent to the power method on the matrix $(1 - \eta w_t^\top A w_t) I_d + \eta A$ at each step, but Oja’s algorithm is equivalent to the power method on the matrix $I + \eta A$. Moreover, $\eta$ should be less than $1/\lambda_1$ for Krasulina’s method to make $(1 - \eta w_t^\top A w_t) I_d + \eta A$ positive, but there is no restriction for Oja’s algorithm.

Appendix B. Riemannian optimization

We now discuss basic notations of optimization on the Riemannian manifold. Given a manifold $\mathcal{M}$ and $x \in \mathcal{M}$, the tangent space is denoted by $T_x \mathcal{M}$, the inner product on $T_x \mathcal{M}$ is defined as $\langle \cdot, \cdot \rangle_x$. The retraction $\mathcal{R}(x, \eta g)$ along the direction $g \in T_x \mathcal{M}$ is a smooth map from $T_x \mathcal{M}$ to $\mathcal{M}$ such that

$$\mathcal{R}(x, 0) = x, \quad \mathcal{R}'(x, 0) = g, \quad \text{where} \quad \mathcal{R}'(x, 0) = \left. \frac{d}{d\eta} \mathcal{R}(x, \eta) \right|_{\eta=0}.$$  

The Riemannian gradient $\text{grad} f$ is the vector in $T_x \mathcal{M}$ such that

$$\langle \text{grad} f(x), g \rangle_x = \nabla f(x)^\top g, \quad \forall g \in T_x \mathcal{M}.$$  

Given a step size $\eta > 0$, the gradient descent method based on the retraction updates the iterates as

$$x_{t+1} = \mathcal{R}(x_t, -\eta \text{grad} f(x)).$$
We refer for readers that Absil et al. (2009) provided more details of optimization on Riemannian manifolds. We end this section with a useful lemma in Riemannian optimization. In particular, Lemma 6 establishes the smoothness property of retraction, which is an essential part of the convergence of first-order methods.

**Lemma 6** (Jiang et al., 2017, Lemma 3.2) If $f$ is a differentiable function and its Euclidean gradient $\nabla f$ is $L$-Lipschitz continuous, then for any $t \geq 0$, there holds that

$$f(\mathcal{R}(x, \eta)) \leq f(\mathcal{R}(x, 0)) + \frac{\eta^2 L}{2} \|\mathcal{R}'(x, 0)\|^2 - \eta \langle \nabla f(\mathcal{R}(x, 0)), \mathcal{R}(x, 0) \rangle_x,$$

where $\mathcal{L} = 2L_2G + L_2^2L$ and $\|\nabla f(x)\| < G$ for all $x$. Moreover, if $\mathcal{R}(x, \eta)$ is QR factorization on the Stiefel manifold, we have $L_1 = 1 + \sqrt{2}/2, L_2 = \sqrt{10}/2$.

**Appendix C. Stochastic Recursive Gradient Algorithm**

This section introduces inexact Stochastic Recursive Gradient Algorithm (iSARAH) (Nguyen et al., 2021) that employs mini-batch gradients instead of full gradients in the outer loop. Given a problem of the form

$$\min_w \left\{ f(w) = \sum_{i=1}^{n} f_i(w) \right\},$$

iSARAH is a stochastic variance reduction method which consists of the outer loop and the inner loop. The outer loop computes the initial estimator of the gradient, while the inner loop estimates the gradient recursively by new samples. Specifically, in an outer loop, given an initial iterate $w_0$ and a large $B_s$ samples $\{i_j^{(0)}\}_{j=1}^{B_s}$ drawn uniform from $[n]$, the initial gradient estimator $v_0$ is

$$v_0 = \frac{1}{B_s} \sum_{j=1}^{B_s} \nabla f_{i_j^{(0)}}(w_0),$$

With fresh $b$ samples $\{i_j^{(t)}\}_{j=1}^{b}$ drawn uniform from $[n]$, SARAH recursively updates the estimation of the gradient:

$$v_t = \frac{1}{b} \sum_{j=1}^{b} \left[ \nabla f_{i_j^{(t)}}(w_t) - \nabla f_{i_j^{(t)}}(w_{t-1}) \right] + v_{t-1},$$

and update $w_{t+1} = w_t - \eta v_t$ with a step size $\eta$ for $t = 1, 2, \ldots, m$. After $m$ iterations of the inner loop, use the last iterate of the inner loop to reset $w_0$ and recompute $v_0$. Letting $\mathcal{F}_t = \sigma(w_0, i_j^{(0)}, i_j^{(1)}, \ldots, i_j^{(t)})$ be the $\sigma$-algebra generated by past information up to time $t$, we note that $v_t$ is a biased estimator of the gradient $\nabla f(w_t)$ conditioned on $\mathcal{F}_{t-1}$, i.e. $\mathbb{E}[v_t|\mathcal{F}_{t-1}] = \nabla f(w_t) - \nabla f(w_{t-1}) + v_{t-1}$, implying following facts:

**Proposition 7** $v_t - \nabla f(w_t)$ is a martingale and the variance is bounded by

$$\mathbb{E}[\|v_m - \nabla f(w_m)\|^2|\mathcal{F}_0] = \|v_0 - \nabla f(w_0)\|^2 + \sum_{t=1}^{m} \mathbb{E}[\|v_t - v_{t-1}\|^2|\mathcal{F}_0] - \sum_{t=1}^{m} \mathbb{E}[(f(w_t) - f(w_{t-1}))^2|\mathcal{F}_0].$$
Nguyen et al. (2021) used a similar argument. For completeness, we provide the proof here.

**Proof** Recall $\mathcal{F}_t = \sigma(w_0, i_j^{(1)}, \ldots, i_j^{(t)})$. We know that $\mathbb{E}[v_t | \mathcal{F}_{t-1}] = \nabla f(w_t) - \nabla f(w_{t-1}) + v_{t-1}$, implying

$$
\mathbb{E}[\|v_t - \nabla f(w_t)\|^2 | \mathcal{F}_t] = \mathbb{E}[\|v_t - v_{t-1}\| + |v_{t-1} - \nabla f(w_{t-1})| + |\nabla f(w_{t-1}) - v_{t-1}|]^2 | \mathcal{F}_t]
$$

$$
= \mathbb{E}[\|v_t - v_{t-1}\|^2 | \mathcal{F}_t] + \|v_{t-1} - \nabla f(w_{t-1})\|^2 - \|\nabla f(w_{t-1}) - v_{t-1}\|^2.
$$

By tower property of conditional expectation, we have

$$
\mathbb{E}[\|v_t - \nabla f(w_t)\|^2 | \mathcal{F}_0] = \mathbb{E}[\|v_t - v_{t-1}\|^2 | \mathcal{F}_0] + \mathbb{E}[\|v_{t-1} - \nabla f(w_{t-1})\|^2 | \mathcal{F}_0] - \mathbb{E}[\|\nabla f(w_{t-1}) - v_{t-1}\|^2 | \mathcal{F}_0].
$$

Summing Eq. (12) over $j = 1 \ldots, m$ implies

$$
\mathbb{E}[\|v_m - \nabla f(w_m)\|^2 | \mathcal{F}_0] = \|v_0 - \nabla f(w_0)\|^2 + \sum_{t=1}^{m} \mathbb{E}[\|v_t - v_{t-1}\|^2 | \mathcal{F}_0] - \sum_{t=1}^{m} \mathbb{E}[\|\nabla f(w_{t-1}) - v_{t-1}\|^2 | \mathcal{F}_0],
$$

which completes the proof.

**Appendix D. Proof of Theorem 1**

1) Define $g_t := A_n w_t$, $P_t := I_d - w_t w_t^\top$, $\theta_t = \theta(w_t, U_k)$ and $\psi_t = \psi(w_t, U_k)$. Let $u$ be an arbitrary unit vector such that $u \in \text{span}(U_k)$.

First, we derive that

$$
\Delta_k \sin^2 \theta(w, U_k) \leq \lambda_1 - w^\top A_n w \leq (\lambda_1 - \lambda_d) \sin^2 (w, U_k),
$$

and

$$
\|P_t g_t\| \leq 2\lambda_1 \sin \theta(w_t, U_k).
$$

Then we have

$$
|u^\top (w_t + \eta P_t g_t + \eta P_t \varepsilon_t)|^2 = |u^\top (w_t + \eta P_t g_t) + \eta u^\top P_t \varepsilon_t|^2
$$

$$
\geq (u^\top (w_t + \eta P_t g_t)) \left((u^\top (w_t + \eta P_t g_t)) + 2\eta u^\top P_t \varepsilon_t\right)
$$

$$
\geq (u^\top (w_t + \eta P_t g_t)) \left((u^\top (w_t + \eta P_t g_t)) - 2\eta \|u^\top P_t \| \| \varepsilon_t \|\right)
$$

$$
\geq (1 + \eta \Delta_k \sin^2 \theta(w_t, u))|u^\top w_t|^2 \left(1 - \frac{2\eta \|u^\top P_t \| \| \varepsilon_t \|}{|u^\top w_t|}\right),
$$

where we have used Cauchy Schwarz inequality (C-S ineq), and

$$
\|w_t + \eta P_t g_t + \eta P_t \varepsilon_t\|^2 = 1 + \eta^2 \|P_t g_t + P_t \varepsilon_t\|^2
$$

$$
\leq 1 + 2\eta^2 \|P_t g_t\|^2 + \|P_t \varepsilon_t\|^2
$$

$$
\leq 1 + 8\eta^2 \lambda_1^2 \sin^2 \theta(w_t, u) + 2\eta^2 \| \varepsilon_t \|^2.
$$
Recall two logarithm inequalities:

\[ \frac{x}{1 + x} \leq \log(1 + x) \leq x \quad \text{for} \quad x > -1, \]  

(17)

and

\[ \frac{x}{1 - \log(1 - x)} \geq \frac{1}{1 - \log(1 - x)} \quad \text{for any} \quad x \in (0, 1). \]  

(18)

Then combining two inequality Eq. (15) and Eq. (16) yields

\[ \psi(w_{t+1}, u) \leq \psi(w_t, u) - \log(1 + \eta \Delta_k \sin^2 \theta(w_t, u)) - \log \left( 1 - \frac{2\eta \|u^T P_t \| \|\varepsilon_t\|}{\|u^T w_t\|} \right) 
\]

\[ + \log(1 + 8\eta^2 \lambda^2_t \sin^2 \theta(w_t, u) + 2\eta^2 \|\varepsilon_t\|^2) \]  

(19)

Taking the minimum with respect to \( u \) over the span\((U_k)\) on both sides of Eq. (19) implies

\[ \psi_{t+1} \leq \psi_t - \eta \left( \frac{\Delta_k}{1 + \eta \Delta_k} - 8\eta \lambda^2_t \right) \sin^2 \theta_t - \log \left( 1 - \frac{2\eta \|\sin \theta(w_t, u)\| \|\varepsilon_t\|}{\|\cos \theta(w_t, u)\|} \right) + 2\eta^2 \|\varepsilon_t\|^2. \]  

Now, we would like to show that \( \psi_t \leq \psi_0 \) for all \( t \) by carefully selecting parameters by

the induction argument under conditions that

\[ \|\varepsilon_t\|^2 \leq \rho^2 \sin^2 \theta_0, \quad \cos^2 \theta_0 \geq \gamma. \]  

(20)

First, for \( t = 1 \), the conditions Eq. (20) implies

\[ \psi_1 \leq \psi_0 - \eta \left( \frac{\Delta_k}{1 + \eta \Delta_k} - 8\eta \lambda^2_t \right) \sin^2 \theta_0 - \log \left( 1 - \frac{2\eta \|\sin \theta_0\| \|\varepsilon_0\|}{\|\cos \theta_0\|} \right) + 2\eta^2 \|\varepsilon_0\|^2 \]

\[ \leq \psi_0 - \eta \left( \frac{\Delta_k}{1 + \eta \Delta_k} - 8\eta \lambda^2_t \right) \sin^2 \theta_0 - \log \left( 1 - \frac{2\eta \rho}{\gamma} \right) + 2\eta^2 \|\varepsilon_0\|^2 \]

\[ \leq \psi_0 - \eta \left( \frac{\Delta_k}{1 + \eta \Delta_k} - 8\eta \lambda^2_t \right) \sin^2 \theta_0 + \frac{2\eta \rho}{\gamma} - 2\eta^2 \rho^2 + 2\eta^2 \|\varepsilon_0\|^2 \]

\[ \leq \psi_0 - \left( \frac{\eta \Delta_k}{2 + 2\eta \Delta_k} - \frac{2\eta \rho}{\gamma - 2\eta \rho} - 2\eta^2 \rho^2 \right) \sin^2 \theta_0, \]

provided that

\[ \gamma - 2\eta \rho > 0 \]  

(21)

and

\[ \eta \Delta_k / (2 + 2\eta \Delta_k) - 8\eta^2 \lambda^2_t > 0. \]  

(22)

Thus, the inequality

\[ \frac{\eta \Delta_k}{2 + 2\eta \Delta_k} - \frac{2\eta \rho}{\gamma - 2\eta \rho} - 2\eta^2 \rho^2 > 0, \]  

(23)
yields $\psi_1 \leq \psi_0$. This completes the case $t = 1$. Suppose it is true that $\psi_t \leq \psi_0$. By the hypothesis, it follows that

$$\sin^2 \theta_t = \frac{\psi_t \sin^2 \theta_t}{-\log(1 - \sin^2 \theta_t)} \geq \frac{\psi_t}{-\log(1 - \sin^2 \theta_t)} \geq \frac{\psi_t}{1 + \psi_0}. \quad (24)$$

Then, since $\cos^2 \theta_t = \exp(-\psi_t)$ and $\sin^2 \theta_t \leq \psi_t$, we get

$$\cos^2 \theta_t \geq \cos^2 \theta_0, \quad \sin^2 \theta_t < \psi_t < \psi_0 \leq (1 + \psi_0) \sin^2 \theta_0, \quad (25)$$

where the last inequality follows by Eq. (24). Using Eq. (25) and the conditions Eq. (20), we have

$$\psi_{t+1} \leq \psi_t - \eta \left( \frac{\Delta_k}{1 + \eta \Delta_k} - 8\eta \lambda_1^2 \right) \sin^2 \theta_t - \log \left( 1 - \frac{2\eta \|\varepsilon_t\|}{\cos \theta_t} \right) + 2\eta^2 \|\varepsilon_t\|^2$$

$$\leq \left( 1 - \frac{\eta \Delta_k}{2 + 2\eta \Delta_k} \frac{1}{1 + \psi_0} \right) \psi_t + \left( \frac{2\eta \rho(1 + \psi_0)}{\gamma - 2\eta \rho} + 2\eta^2 \rho^2 \right) \sin^2 \theta_0, \quad (26)$$

provided that $\gamma - 2\eta \rho > 0$ and

$$\eta \Delta_k/(2 + 2\eta \Delta_k) - 8\eta^2 \lambda_1^2 > 0. \quad (27)$$

Therefore, $\psi_{t+1} < \psi_0$ follows if it holds that

$$\frac{\eta \Delta_k}{2 + 2\eta \Delta_k} \frac{1}{1 + \psi_0} - \frac{2\eta \rho(1 + \psi_0)}{\gamma - 2\eta \rho} - 2\eta^2 \rho^2 > 0. \quad (28)$$

Hence, we prove that $\psi_t \leq \psi_0$ since our assumption on $\eta$ Eq. (4) implies Eq. (21), Eq. (22), Eq. (23), Eq. (27), Eq. (28).

For the second statement, let $r = \eta \Delta_k/(2 + 2\eta \Delta_k)$. Since Eq. (4) implies that

$$\frac{\eta \Delta_k}{4 + 4\eta \Delta_k} \frac{\beta}{1 + \psi_0} \geq \frac{2\eta \rho(1 + \psi_0)}{\gamma - 2\eta \rho} + 2\eta^2 \rho^2, \quad (29)$$

we obtain by Eq. (26) and Eq. (29)

$$\psi_{t+1} \leq (1 - r)\psi_t + \frac{r \beta}{2} \psi_0 \leq (1 - r)^t \psi_0 + \frac{r \beta}{2} \psi_0 \sum_{i=0}^{\infty} (1 - r)^i \leq \left( (1 - r)^t + \frac{\beta}{2} \right) \psi_0,$$

and the theorem for general $d$ follows.

2) For the special case $d = 2$, let $\theta = \theta(w, u_1)$ and $\theta' = \theta(w', u_1)$. We use $\varepsilon_t = \varepsilon$, $w_t = w$ and $w_{t+1} = w'$ for simplify notations. Assume that $w = s_1 u_1 + s_2 u_2$. Then, $A_n w = \lambda_1 s_1 e_1 + \lambda_2 s_1 e_1$, which implies

$$w' = w + \eta P(A_n w + \varepsilon)$$

$$= w + \eta A_n w - \eta w^\top A_n w + \eta P \varepsilon$$

$$= \sum_{i=1}^{2} [s_i (1 + \eta \lambda_i - \eta w^\top A w) + \eta u_i^\top P \varepsilon] u_i.$$
Thus, we have
$$1 + \eta \lambda_i - \eta w^\top Aw > 0.$$  
(30)

Moreover, since $w^\top A_n w = \lambda_1 s_1^2 + \lambda_2 s_2^2$ and $s_1^2 + s_2^2 = 1$, we know
$$\lambda_1 - w^\top Aw = (\lambda_1 - \lambda_1 s_1^2) - \lambda_2 s_2^2 = (\lambda_1 - \lambda_2) s_2^2 = (\lambda_1 - \lambda_2) \sin^2 \theta.$$  
(31)

Since $\|\varepsilon\| < \rho_1 (\lambda_1 - \lambda_2) |\cos \theta| = \rho_1 (\lambda_1 - \lambda_2) |s_1|$, we obtain
$$|u_1^\top P \varepsilon| / |s_1| \leq \rho_1 (\lambda_1 - \lambda_2).$$  
(32)

Then, we have
$$|\tan \theta'| = \frac{s_2 (1 + \eta \lambda_2 - \eta w^\top Aw) + \eta u_2^\top P \varepsilon|}{s_1 (1 + \eta \lambda_1 - \eta w^\top Aw) + \eta u_1^\top P \varepsilon|} \leq \frac{s_2 (1 + \eta \lambda_2 - \eta w^\top Aw) + \eta u_2^\top P \varepsilon|}{s_1 (1 + \eta \lambda_1 - \eta w^\top Aw) - \eta u_1^\top P \varepsilon|}$$
$$Eq. \ (30) = \frac{s_2 (1 + \eta \lambda_2 - \eta w^\top Aw) + \eta u_2^\top P \varepsilon|}{s_1 (1 + \eta \lambda_1 - \eta w^\top Aw) - \eta u_1^\top P \varepsilon|} = \frac{1 + \eta \lambda_2 - \eta w^\top Aw}{s_1 (1 + \eta \lambda_1 - \eta w^\top Aw) - \eta u_1^\top P \varepsilon|/|s_1|} + \frac{1}{|s_1|} \left( 1 - \eta |u_1^\top P \varepsilon| / |s_1| \right)$$
$$Eq. \ (31) = |\tan \theta| \left( 1 - \frac{\eta (\lambda_1 - \lambda_2) - \eta |u_1^\top P \varepsilon| / |s_1|}{\eta |u_2^\top P \varepsilon| / |s_1|} \right) + \frac{\eta |u_2^\top P \varepsilon|}{|s_1|}$$
$$\leq |\tan \theta| \left( 1 - \frac{(\eta - \rho_1) (\lambda_1 - \lambda_2)}{1 + (\eta - \rho_1) (\lambda_1 - \lambda_2)} \right) + \frac{\eta |\varepsilon|}{|s_1|}$$
$$Eq. \ (32) \leq \frac{|\tan \theta|}{1 + (\eta - \rho_1) (\lambda_1 - \lambda_2)} + \frac{\eta |\varepsilon|}{1 + (\eta - \rho_1) (\lambda_1 - \lambda_2)}$$
$$Eq. \ (33) \leq \frac{1 + \eta \rho_2 (\lambda_1 - \lambda_2)}{1 + (\eta - \rho_1) (\lambda_1 - \lambda_2)} |\tan \theta|,$$

provided that
$$|\cos \theta| > \gamma, \quad \|\varepsilon\| \leq \rho_2 (\lambda_1 - \lambda_2) |\sin \theta|. \quad (33)$$

Thus, we have $|\tan \theta'| \leq |\tan \theta|$ since we assume $\eta \rho_2 \leq \gamma (\eta - \rho_1)$. As a result, $|\cos \theta'| \geq \gamma$, by the same induction argument for general $d > 2$, we conclude $|\tan \theta_t| \leq |\tan \theta_0|$ and
$$|\tan \theta_{t+1}| \leq \left( 1 - \frac{(\eta - \rho_1) (\lambda_1 - \lambda_2)}{1 + (\eta - \rho_1) (\lambda_1 - \lambda_2)} \right) |\tan \theta_t| + \frac{\eta \rho_2 (\lambda_1 - \lambda_2) / \gamma}{1 + (\eta - \rho_1) (\lambda_1 - \lambda_2)} |\tan \theta_0|.$$  

Let $r = (\eta - \rho_1) (\lambda_1 - \lambda_2)/(1 + (\eta - \rho_1) (\lambda_1 - \lambda_2))$. Using the condition
$$\frac{\eta \rho_2 (\lambda_1 - \lambda_2) / \gamma}{1 + (\eta - \rho_1) (\lambda_1 - \lambda_2)} \leq \frac{\beta}{2} \frac{(\eta - \rho_1) (\lambda_1 - \lambda_2)}{1 + (\eta - \rho_1) (\lambda_1 - \lambda_2)} = \frac{\beta r}{2}$$

with $\beta > 1$, we have $|\tan \theta_{t+1}| \leq \beta |\tan \theta_t|$.
we get
\[ |\tan \theta_{t+1}| \leq (1 - r)|\tan \theta_t| + \frac{r \beta}{2} |\tan \theta_0| \]
\[ \leq (1 - r)^t |\tan \theta_0| + \sum_{k=1}^{t} (1 - r)^{t-k} \frac{r \beta}{2} |\tan \theta_0| \]
\[ \leq (1 - r)^t |\tan \theta_0| + \frac{\beta}{2} |\tan \theta_0|, \]
which completes the proof.

**Appendix E. Proof of Proposition 3**

1) We first show the first part of Proposition 3 for general $d$. Let $f(w_t) = \frac{1}{2} w_t^\top A_n w_t$. Then $f$ is $K$-smooth by the condition
\[ \max \{\|a_t\|^2, \|a_t q_t^\top - A_n\|\} \leq K. \quad (34) \]

Recall that $\mathcal{R}(w_{t-1}, \eta P_{t-1} v_{t-1}) = w_t$, $\mathcal{R}'(w_{t-1}, 0) = P_{t-1} v_{t-1}$, and $P_t \nabla f(w_t)$ is Riemannian gradient (Absil et al., 2009, chapter 4).

Define $E_0 = \mathbb{E}[\cdot|\mathcal{F}_0]$. By Lemma 6 and the fact that $r^\top q = \frac{1}{2} [||r||^2 + ||q||^2 + ||r - q||^2]$, we get
\[
E_0[f(w_t)] \leq E_0[f(w_{t-1})] + \frac{\eta^2 \bar{L}}{2} E_0[||P_{t-1} v_{t-1}||^2 - \eta E_0(P_{t-1} \nabla f(w_{t-1}), P_{t-1} v_{t-1})w_{t-1}]
\]
\[= E_0[f(w_{t-1})] - \left(\eta - \frac{\eta^2 \bar{L}}{2}\right) E_0[||P_{t-1} v_{t-1}||^2 - \eta E_0||P_{t-1} \nabla f(w_{t-1})||^2
\]
\[+ \eta E_0||P_{t-1} \nabla f(w_{t-1}) - P_{t-1} v_{t-1}||^2, \]
which implies
\[ E_0[f(w_t)] \leq E_0[f(w_{t-1})] - \left(\eta - \frac{\eta^2 \bar{L}}{2}\right) E_0[||P_{t-1} v_{t-1}||^2 + \eta E_0||v_{t-1} - \nabla f(w_{t-1})||^2. \quad (35) \]

Moreover, we have
\[ E_0[||v_t - v_{t-1}||^2] = \mathbb{E}_0\left[\frac{1}{b} \sum_{j=1}^{b} \left(\nabla f_{j;i}(w_t) - \nabla f_{j;i}(w_{t-1})\right)^2\right] \leq \frac{K^2}{b} \mathbb{E}_0[||w'_t - w_{t-1}||^2] \leq \frac{K^2 \eta^2}{b} \mathbb{E}_0[||P_{t-1} v_{t-1}||^2], \quad (36) \]
where we have used the condition Eq. (34) and the updating rule
\[ v_t = \frac{1}{b} \sum_{j=1}^{b} \left[\nabla f_{j;i}(w_t) - \nabla f_{j;i}(w_{t-1})\right] + v_{t-1}, \]
and \( w'_t \) is the vector before projection and the last two inequality is due to the property of projection. Combining Lemma 7 and Eq. (36) yields
\[
\sum_{t=1}^{m} \mathbb{E}_0 \| v_{t-1} - \nabla f(w_{t-1}) \|^2 \leq \sum_{t=1}^{m} \sum_{k=1}^{t} \mathbb{E}_0 \| v_k - v_{k-1} \|^2 + m \| v_0 - \nabla f(w_0) \|^2 \\
\leq \frac{m \eta^2 K^2}{b} \sum_{t=1}^{m} \mathbb{E}_0 \| P_{t-1} v_{t-1} \|^2 + m \| v_0 - \nabla f(w_0) \|^2. \tag{37}
\]

By telescoping Eq. (35), we have
\[
\left( \eta - \frac{\eta^2 \hat{L}}{2} \right) \sum_{t=1}^{m} \mathbb{E}_0 \| P_{t-1} v_{t-1} \|^2 \\
\leq \mathbb{E}_0 [ f(w_0) - f(w_{m+1}) ] + \eta \sum_{t=1}^{m} \mathbb{E}_0 \| v_{t-1} - \nabla f(w_{t-1}) \|^2 \\
\leq f(w_0) - f(w^*) + \eta m \mathbb{E}_0 \| v_0 - \nabla f(w_0) \|^2 + \frac{m \eta^3 K^2}{b} \sum_{t=1}^{m} \mathbb{E}_0 \| P_{t-1} v_{t-1} \|^2,
\]
where \( w^* = \arg \min_{w \in \mathcal{M}} f(w) \) and the last inequality follows by Eq. (37). Therefore, since
\[
f(w_0) - f(w^*) = \lambda_1 - w_0^\top A_n w_0 \leq \lambda_1 (1 - (w^*_0 u_1)^2) = \lambda_1 \sin \theta_0,
\]
we get
\[
\left( \eta - \frac{\eta^2 \hat{L}}{2} - \frac{m \eta^3 K^2}{b} \right) \sum_{t=1}^{m} \mathbb{E} [ \| P_t v_t \|^2 | \mathcal{F}_t ] \leq \lambda_1 \sin^2 \theta_0 + \eta m \| v_0 - A_n w_0 \|, \tag{38}
\]
and complete the first statement.

Now we show the second statement. By Jensen’s inequality for conditional expectation and the fact of martingale \( \mathbb{E}[v_t - A_n w_t | \mathcal{F}_{t-1}] = v_{t-1} - A_n w_{t-1} \), we have
\[
\mathbb{E} \left[ \frac{\| v_t - A_n w_t \|^2}{\sin^2 \theta_0} \right] \geq \mathbb{E} \left[ \frac{\| v_{t-1} - A_n w_{t-1} \|^2}{\sin^2 \theta_0} \right] = \frac{\| v_{t-1} - A_n w_{t-1} \|^2}{\sin^2 \theta_0}.
\]
Thus, \( \| v_t - A_n w_t \|^2 / \sin^2 \theta_0 \) is the submartingale. By Doob’s maximal inequality for submartingale (Durrett, 2019, Thm 5.4.2), we have
\[
\mathbb{P} \left\{ \max_{1 \leq t \leq m} \frac{\| v_t - A_n w_t \|^2}{\sin^2 \theta_0} > \rho^2 \right\} \leq \frac{1}{\rho^2} \mathbb{E} \left[ \frac{\| v_m - A_n w_m \|^2}{\sin^2 \theta_0} \right].
\]

Define \( \tau := 1 - \frac{\eta \hat{L}}{2} - \frac{m \eta^2 K^2}{b} \). By Lemma 7 and the first part of proposition 3, we have
\[
\mathbb{E}_0 \left[ \frac{\| v_m - A_n w_m \|^2}{\sin^2 \theta_0} \right] \leq \frac{1}{\sin^2 \theta_0} \left[ \| v_0 - A_n w_0 \|^2 + \sum_{t=1}^{m} \mathbb{E}_0 \| v_t - v_{t-1} \|^2 \right] \\
\leq \frac{K^2}{B_s \sin^2 \theta_0} + \frac{K^2 \eta^2}{b \sin^2 \theta_0} \sum_{t=1}^{m} \mathbb{E}_0 \| P_{t-1} v_{t-1} \|^2 \\
\leq \frac{K^2}{B_s \sin^2 \theta_0} + \frac{K^2 \eta}{b \tau \sin^2 \theta_0} \left( \lambda_1 \sin^2 \theta_0 + \frac{\eta m K^2}{B_s} \right),
\]
where we have used $v_0 = \frac{1}{B_s} \sum_{j=1}^{B_s} a_{i,j}(a_{i,j})^\top w_0$ in the second inequality. Thus, if we choose $B_s = B/\sin^2\theta_0$, we have

$$\mathbb{P}\left\{ \max_{1 \leq t \leq m} \frac{\|v_t - A_n w_t\|^2}{\sin^2\theta_0} > \rho^2 \right\} \leq \frac{1}{\rho^2} \mathbb{E} \left[ \mathbb{E} \left[ \frac{\|v_m - A_n w_m\|^2}{\sin^2\theta_0} | F_0 \right] \right] \leq \frac{K^2}{\rho^2 B} + \frac{K^2}{\rho^2 B} + \frac{m \eta^2 K^4}{b \tau \rho^2 B}. \tag{39}$$

Therefore, the proposition for general $d$ follows.

2) in the two-dimensional space, by Eq. (31) we have the tight bound

$$f(w_0) - f(w_*) = \lambda_1 - w_0^\top A_n w_0 = (\lambda_1 - \lambda_2) \sin \theta_0.$$ 

Therefore, Eq. (38) and Eq. (39) become

$$\left( \eta - \frac{\eta^2 \tilde{L}}{2} - m \eta^3 \frac{K^2}{b} \right) \sum_{t=1}^{m} \mathbb{E}[\|P_t v_t\|^2 | F_0] \leq \Delta_1 \sin^2\theta_0 + \eta m \|v_0 - A_n w_0\|,$$

and

$$\mathbb{P}\left\{ \max_{1 \leq t \leq m} \frac{\|v_t - A_n w_t\|^2}{\sin^2\theta_0} > \rho^2 \right\} \leq \frac{1}{\rho^2} \mathbb{E} \left[ \mathbb{E} \left[ \frac{\|v_m - A_n w_m\|^2}{\sin^2\theta_0} | F_0 \right] \right] \leq \frac{K^2}{\rho^2 B} + \frac{\Delta_1 K^2 \eta}{\rho^2 b \tau} + \frac{m \eta^2 K^4}{b \tau \rho^2 B}.$$

The proposition follows.

**Appendix F. Proof of Theorem 4**

1) It suffices to prove convergence result for one loop of Algorithm 1. Noting that $v_t = v_t - A w_t$ for iRSRG, we define two events

$$E = \{ \|v_t\|^2 < \rho^2 \sin^2\theta_0 | t = 1, \ldots, m \}$$

and

$$F = \{ \cos \theta_0 > \gamma \}.$$ 

Then we could apply Theorem 1 under events $E$ and $F$. Moreover, Proposition 3 implies the probability of event $E$ can be bounded as

$$\mathbb{P}(E^c) \leq \frac{K^2}{\rho^2 B} + \frac{K^2 \eta}{\rho^2 b \tau} + \frac{m \eta^2 K^4}{b \tau \rho^2 B}.$$ 

For random initialization, we use the following lemmas to bound the probability of event $F$. In particular, Lemma 8 controls the distance between random initialization and the first eigenspace.
Lemma 8 (Xu and Gao, 2018, Lemma 4.7) For a uniformly sampled point \( w_0 \in \text{Grass}(d, k) \) and \( 0 < \gamma < 1 \), we have that \( \cos^2 \theta(U_k, w_0) \geq \gamma \) with probability at least

\[
1 - p_k(\gamma) = \frac{\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{d-k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+1}{2}\right)} (\sin(\gamma^{-1}(\gamma^{1/(2k)}))^{k(d-k)} \right)_{2F1} \left( \frac{d-k}{2}, \frac{1}{2}, \frac{d+1}{2}; I_k \sin^2(\gamma^{-1}(\gamma^{1/(2k)})) \right),
\]

where \( 2F1 \) is the Gaussian hypergeometric function of matrix argument.

If we pick \( m = \frac{1}{r} \log \frac{2}{\beta} \), then we get

\[
\tilde{\psi}_1 \leq \left( (1-r)^m + \frac{\beta}{2} \right) \tilde{\psi}_0 \leq \left( \exp(-rm) + \frac{\beta}{2} \right) \tilde{\psi}_0 = \beta \tilde{\psi}_0.
\]

The final conclusion simply follows by the union bound.

2) For \( d = 2 \), it suffices to control the probability of the following event

\[ E = \{ \|\varepsilon_t\| \leq \min\{\rho_1\Delta_1|\cos \theta_0|, \rho_2\Delta_1|\sin \theta_0|\} \mid t = 1, \ldots, m \}, \]

where \( \varepsilon_t = v_t - Aw_t \). On the event \( \{ |\cos \theta_0| \geq \gamma \} \), we know

\[ \|\varepsilon\| \leq \min\{\rho_1\Delta_1|\cos \theta_0|, \rho_2\Delta_1|\sin \theta_0|\} \Leftrightarrow \|\varepsilon\| \leq \rho_1\rho_2\gamma\Delta_1|\sin \theta_0|. \]

Hence, it suffice to bound

\[ E' = \{ \|\varepsilon_t\| \leq \rho_1\rho_2\gamma\Delta_1|\sin \theta_0| \mid t = 1, \ldots, m \}. \]

By Proposition 3, we

\[
\mathbb{P}(E') \leq \frac{K^2}{(\rho_1\rho_2\gamma\Delta_1)^2B} + \frac{\Delta_1K^2\eta}{(\rho_1\rho_2\gamma\Delta_1)^2b\tau} + \frac{m\eta^2K^4}{b\tau(\rho_1\rho_2\gamma\Delta_1)^2B}.
\]

Appendix G. Proof of Corollary 5

1) Note that \( \Delta_k \leq 1 \) and \( \beta = 0.5 \). Let \( \eta = c_\eta \Delta_k \), and \( \rho = c_\rho \Delta_k \) for \( c_\eta \leq 1 \). Since \( \eta \leq 1 \), we have

\[
\gamma \rho - 2\eta \rho^2 \geq \frac{\gamma \rho}{2} \Leftrightarrow \rho \leq \frac{1}{4\eta} \Leftrightarrow \rho \leq \frac{1}{4},
\]

yielding that

\[
\frac{1 + \psi_0}{\gamma \rho - 2\eta \rho^2} \leq \frac{\Delta_k}{4\rho^2(8 + 8\eta \Delta_k)} \frac{1}{1 + \psi_0} \Leftrightarrow 64(1 + \psi_0)^2 \rho^2 \leq \Delta_k(\gamma \rho - 2\eta \rho^2)
\]

\[
\Leftrightarrow 64(1 + \psi_0)^2 \rho^2 \leq \frac{\Delta_k \gamma \rho}{2}
\]

\[
\Leftrightarrow \rho \leq \frac{\Delta_k \gamma}{128(1 + \psi_0)^2}.
\]
Hence, we know that

\[ c_\eta \leq \min \left\{ \frac{\gamma}{2}, \frac{1}{32\lambda_1}, \frac{1}{16(1 + \psi_0)} \right\}, \quad c_\rho \leq \min \left\{ \frac{1}{4}, \frac{\gamma}{128(1 + \psi_0)^2} \right\} \]

implies

\[ \eta < \frac{\gamma}{2\rho} \land \frac{\Delta_k}{16\lambda_1(1 + \eta\Delta_k)} \land \frac{\Delta_k}{2\rho^2(8 + 8\eta\Delta_k)(1 + \psi_0)} \land \frac{1}{\gamma \rho - 2\eta\rho^2} \]

Moreover, since \( \eta\Delta_k \leq 1 \) and \( m \geq \log(4)/r \), we can choose

\[ m = \left\lceil \frac{3\log(4)}{c_\eta \Delta_k^2} \right\rceil \geq \left\lceil \frac{\log(4)}{r} \right\rceil = \left\lceil \frac{(2 + 2\eta\Delta_k)\log(4)}{\eta\Delta_k} \right\rceil \]

and set \( b = \left\lceil \max\{\lambda_1K^2, K^2\rho\} \right\rceil, B = \left\lceil c_B^\kappa \right\rceil \).

Note that

\[ 1 - \eta\hat{L} - m\eta^2K^2 \geq \frac{1}{2} \iff 1 - c_\eta - \eta\hat{L}\Delta_k - 3\log(4)c_\eta\rho > 0 \iff c_\eta < \frac{1}{2(\hat{L} + 3\log(4))}. \]

Since \( 1 - p_k(\gamma) - \frac{99}{100}p_k(\gamma) \geq 0.5 \), we get

\[ 1 - M \left( \frac{K^2}{\rho^2B} + \frac{\lambda_1 K^2\eta}{\rho^2\tau^2} + \frac{m\eta^2K^4}{\rho^2\tau^2B} \right) - p_k(\gamma) \geq \frac{99}{100}p_k(\gamma) \]

\[ \iff \frac{K^2}{\rho^2B} + \frac{\lambda_1 K^2\eta}{\rho^2\tau^2} + \frac{m\eta^2K^4}{\rho^2\tau^2B} \leq \frac{1}{2M} \]

\[ \iff \frac{1}{c_B} + 2\frac{c_\eta}{c_\rho} + 6\log(4)c_\eta c_\rho \leq \frac{1}{2M} \]

\[ \iff c_B \geq 36\log(4)M, \quad c_\eta \leq \frac{1}{6c_\rho M}. \]

In sum, to meet the requirement of (2) in Theorem 4, we need that

\[ c_B \geq 36\log(4)M, \quad c_\eta \leq \min \left\{ \frac{\gamma}{2}, \frac{1}{32\lambda_1}, \frac{1}{16(1 + \psi_0)}, \frac{1}{6c_\rho M} \right\}, \quad c_\rho \leq \min \left\{ \frac{1}{4}, \frac{\gamma}{128(1 + \psi_0)^2} \right\}. \]

Since \( M = \log(\psi_0/\text{tol})/\log(2) \), the number of samples that iRSRG uses to achieve tol-accuracy is

\[ M \left( \frac{B}{\text{tol}} + mb \right) = \mathcal{O} \left( \frac{\log(1/\text{tol})}{\Delta_k^2 \text{tol}} + \frac{1}{\Delta_k} \log \left( \frac{1}{\text{tol}} \right) \right), \]

which completes the proof.

2) Note that \( \beta = 0.5 \). Let \( \eta = c_\eta/\lambda_1 \) and \( \rho_1 = \eta/2 \) such that \( c_\eta < 1 \). Then we know

\[ \gamma(\eta - \rho_1) - 4\rho_2\eta \geq 0 \iff \rho_2 \leq \frac{1}{8\gamma}. \]
Moreover, since \( \Delta_1 \leq \frac{1}{\lambda_1} \) and \( m \geq \log(4)/r \), we can choose \( m \) to be

\[
m = \left\lceil \frac{3 \log(4)}{\eta \Delta_1} \right\rceil \geq \left\lfloor \frac{\log(4)}{r} \right\rfloor = \left\lfloor \frac{\log(4)(2 + \eta \Delta_1)}{\eta \Delta_1} \right\rfloor.
\]

Then we set \( b = \left\lceil K^2 \frac{1}{\lambda_1 \Delta_1} \right\rceil \), \( B = \left\lceil c_B \frac{K^2}{\rho_1^2 \rho_2^2 \Delta_1} \right\rceil \). We have

\[
1 - \frac{\eta \hat{L}}{2} - \frac{m \eta^2 K^2}{b} > \frac{1}{2} \iff \frac{1}{2} - \frac{\hat{L} c_\eta}{2 \lambda_1} - 3 \log(4) c_\eta > 0 \iff c_\eta < \frac{\lambda_1}{\hat{L} + 6 \log(4) \lambda_1}.
\]

Since we choose \( \gamma \) such that \( 1 - p_k(\gamma) - \frac{99}{100} p_k(\gamma) \geq 0.5 \), we get

\[
1 - M \left( \frac{K^2}{\gamma^2 \rho_1^2 \rho_2^2 \Delta_1^2 B} + \frac{K^2 \eta}{\gamma^2 \rho_1^2 \rho_2^2 \Delta_1 b r} + \frac{m \eta^2 K^4}{b r \gamma^2 \rho_1^2 \rho_2^2 \Delta_1^2 B} \right) - p_k(\gamma) \geq \frac{99}{100} p_k(\gamma)
\]

\[
\iff \frac{1}{c_B} + \frac{2 c_\eta}{\gamma^2 \rho_1^2 \rho_2^2} + 3 \log(4) \frac{c_\eta}{c_B} \leq \frac{1}{2 M}
\]

\[
\iff c_B \geq \frac{36 \log(4) M}{\lambda_1}, \quad c_\eta \leq \frac{1}{6 M \gamma^2 \rho_1^2 \rho_2^2}.
\]

In sum, to meet the assumptions in Theorem 4, we need that

\[
\rho_1 = \frac{\eta}{2}, \quad \rho_2 \leq \frac{1}{8 \gamma}, \quad c_B \geq 36 \log(4) M, \quad c_\eta \leq \min \left\{ \frac{1}{6 M \gamma^2 \rho_1^2 \rho_2^2}, \frac{\lambda_1}{\hat{L} + 6 \log(4) \lambda_1} \right\}.
\]

Since \( M = \log(\psi_0/\text{tol})/\log(1/\beta) \), the number of samples that iRSRG uses to achieve \( \text{tol}-\)accuracy in the two-dimensional space is

\[
M \left( \frac{B}{\text{tol}} + m b \right) = \mathcal{O} \left( \left( \frac{\log(1/\text{tol})}{\Delta_k^2 \text{tol}} + \frac{1}{\Delta_k} \right) \log \left( \frac{1}{\text{tol}} \right) \right).
\]

This completes the proof.