A Quantitative Analysis of the Effect of Batch Normalization on Gradient Descent Appendix

A. Batch and more general normalization on general objective functions

Here we consider the generalized versions of batch normalization on general problems, including but not limited to deep neural networks. Consider a smooth loss function $J_0(w_1, ..., w_m)$ and its normalized version $J(\gamma_1, ..., \gamma_m, w_1, ..., w_m)$,

$$J(\gamma_1, ..., \gamma_m, w_1, ..., w_m) = J_0\big(\gamma_1 \tfrac{w_1}{\|w_1\|_{S_1}}, ..., \gamma_m \tfrac{w_m}{\|w_m\|_{S_m}}\big), w_i \neq 0, i = 1, ..., m. \tag{18}$$

Here the normalizing matrices $S_i, i = 1, ..., m$, are assumed to be positive definite and S_i does not depend on w_i and γ_i (it could depend on w_j or $\gamma_j, j < i$). For neural networks, choosing $S_i = I$ as the identity matrix, one gets the weight normalization (Salimans & Kingma, 2016). Choosing S_i as the covariance matrix Σ_i of ith layer output z_i, one gets batch normalization. When the covariance matrix is degenerate, one can set $S_i = \Sigma_i + S_0$ with S_0 being small but positive definite, e.g. $S_0 = 0.001I$.

It is obvious that the normalization changes the landscape of the original loss function J_0, such as introducing new stationary points which are not stationary points of J_0. However, we will show the newly introduced stationary points are strict saddle points and hence can be avoid by many optimization schemes (Lee et al., 2016; Panageas & Piliouras, 2017).

A.1. Normalization only introduces strict saddles

Let us begin with a simple case where $m = 1$ in Eq. (18), i.e. $J(\gamma, w; S) = J_0\big(\gamma \frac{w}{\|w\|_S}\big)$. In this case, the gradients of J are

$$\frac{\partial J}{\partial \gamma} = \nabla J_0\big(\gamma \tfrac{w}{\|w\|_S}\big)^T \frac{w}{\|w\|_S}, \tag{19}$$

$$\frac{\partial J}{\partial w} = \frac{\gamma}{\|w\|_S}\big(I - \frac{Sww^T}{\|w\|_S^2}\big)\nabla J_0\big(\gamma \tfrac{w}{\|w\|_S}\big). \tag{20}$$

The stationary points (γ, w) of J can be grouped into two parts:

(1) $\tilde{w} := \frac{\gamma w}{\|w\|_S}$ is a stationary point of J_0. In this case, $\gamma = \pm\|\tilde{w}\|_S$.

(2) $\tilde{w}$ is not a stationary point of J_0. In this case, $\gamma = 0, w^T \nabla J_0(\tilde{w}) = 0$.

The stationary points in (2) are ones introduced by normalization, giving the Hessian matrix

$$A_1 := \begin{pmatrix} \frac{\partial^2 J}{\partial \gamma^2} & \frac{\partial^2 J}{\partial \gamma \partial w} \\ \frac{\partial^2 J}{\partial w \partial \gamma} & \frac{\partial^2 J}{\partial w^2} \end{pmatrix} = \begin{pmatrix} \frac{w^T(\nabla^2 J_0(\tilde{w}))w}{\|w\|_S^2} & \frac{1}{\|w\|_S^2}(\nabla J_0(\tilde{w}))^T \\ \frac{1}{\|w\|_S^2}\nabla J_0(\tilde{w}) & 0 \end{pmatrix}. \tag{21}$$

Since $\nabla J_0(\tilde{w}) \neq 0$, the rank of A_1 is 2. In fact, the nonzero eigenvalues of A_1 are:

$$\frac{a \pm \sqrt{a^2 + 4\|b\|^2}}{2},$$

where $a = \frac{w^T(\nabla^2 J_0)w}{\|w\|_S^2}, b = \frac{1}{\|w\|_S^2}\nabla J_0$. Therefore A_1 has a negative eigenvalue, and (γ, w) is a strict saddle point.

Let us now consider the case of $m > 1$. The normalization-introduced stationary points satisfy $\gamma_i = 0, w_i^T \nabla J_0(\tilde{w}_i) = 0$. The Hessian matrix A at these points always has negative eigenvalues because it has a principal minor like A_1 in Eq. (21). Thus we have the following lemma:

Lemma A.1. *If $(\gamma_1, ..., \gamma_m, w_1, ..., w_m)$ is a stationary point of J but $(\frac{\gamma_1 w}{\|w_1\|_{S_1}}, ..., \frac{\gamma_m w_m}{\|w_m\|_{S_m}})$ is not a stationary point of J_0, then $(\gamma_1, ..., \gamma_m, w_1, ..., w_m)$ is a strict saddle point of J.*

A.2. Scaling property and increasing norm of w_i

When using gradient descent to minimize the loss function (18), we need to specify the numerical parameters including the initial values of γ_i and w_i, which denoted by Γ_0 and W_0 respectively, and the step size for them, denoted by ε_γ and ε. For simplicity, we use the same ε_γ for all γ_i and the same ε for w_i. Due to the fact that the scale of w_i does not effect the loss, we immediately have the scaling properties on the set of numerical parameters, or a *configuration* $\{\Gamma_0, W_0, \varepsilon_\gamma, \varepsilon\}$.

Definition A.2 (Equivalent configuration)**.** *Two configurations,* $\{\Gamma_0, W_0, \varepsilon_\gamma, \varepsilon\}$ *and* $\{\Gamma_0', W_0', \varepsilon_\gamma', \varepsilon'\}$*, are said to be equivalent if for iterates* $\{\Gamma_k, W_k\}$*,* $\{\Gamma_k', W_k'\}$ *following these configurations respectively, there is an invertible linear transformation* T *and a nonzero constant* t *such that* $W_k' = TW_k, \Gamma_k' = t\Gamma_k$ *for all* k*.*

It is easy to check the gradient descent on normalized loss function (18) has the following scaling property.

Proposition A.3 (Scaling property)**.** *For any* $r \neq 0$*, the configurations* $\{\Gamma_0, W_0, \varepsilon_\gamma, \varepsilon\}$ *and* $\{\Gamma_0, rW_0, \varepsilon_\gamma, r^2\varepsilon\}$ *are equivalent.*

Proof. Gradient descent gives the following iteration:

$$\gamma_{i,k+1} = \gamma_{i,k} - \varepsilon_\gamma \tfrac{\partial J}{\partial \gamma_i}(\Gamma_k, W_k), \tag{22}$$

$$w_{i,k+1} = w_{i,k} - \varepsilon \tfrac{\partial J}{\partial w_i}(\Gamma_k, W_k). \tag{23}$$

It is easy to check that $\frac{\partial J}{\partial (rw_i)} = \frac{1}{r}\frac{\partial J}{\partial w_i}, rw_{i,k+1} = rw_{i,k} - r^2\varepsilon\frac{\partial J}{\partial (rw_i)}(\Gamma_k, W_k)$. Let $\gamma_i = \gamma_i', w_i' = rw_i, \varepsilon_\gamma' = \varepsilon_\gamma, \varepsilon' = r^2\varepsilon$, then we immediately have the equivalence result. □

Another consequence of the invariance of loss functions with respect to the scale of w_i is the orthogonality between w_i and $\frac{\partial J}{\partial w_i}$. In fact, we have $0 = \frac{\partial l}{\partial \|w_i\|} = \frac{w_i}{\|w_i\|} \cdot \frac{\partial J}{\partial w_i}$. As a consequence, we have the following property.

Proposition A.4 (Increaing norm of w_i)**.** *For any configuration* $\{\Gamma_0, W_0, \varepsilon_\gamma, \varepsilon\}$*, the norm of each* w_i *is incresing during gradient descent iteration.*

Proof. According to the orthogonality between w_i and $\frac{\partial J}{\partial w_i}$, we have

$$\|w_{i,k+1}\|^2 = \|w_{i,k}\|^2 + \varepsilon^2 \big\|\tfrac{\partial J}{\partial w_{i,k}}\big\|^2 \geq \|w_{i,k}\|^2, \tag{24}$$

which finishes the proof. □

A.3. Convergence for arbitrary step size

As a consequence of scaling property and the increasing-norm property, we have the following convergence result, which says that convergence for small learning rates implies convergence for arbitrary learning rates for weights.

Theorem A.5 (Convergence of the gradient descent on (18))**.** *If there are two positive constants,* $\varepsilon_\gamma^*, \varepsilon^*$*, such that the gradient descent on* J *converges for any initial value* Γ_0, W_0 *such that* $\|w_{i,0}\| = 1$ *and step size* $\varepsilon_\gamma < \varepsilon_\gamma^*, \varepsilon < \varepsilon^*$*, then the gradient of* w_i *converges for arbitrary step size* $\varepsilon > 0$ *and* $\varepsilon_\gamma < \varepsilon_\gamma^*$*.*

Proof. Firstly, the norm of each $w_{i,k}$ must converge for any step size $\varepsilon > 0$ and $\varepsilon_\gamma < \varepsilon_\gamma^*$. In fact, if $w_{i,k}$ is not bounded, then there is a $k = K$ such that $\frac{\varepsilon}{\|w_{i,K}\|^2} < \varepsilon^*$. Then using the scaling property, one has a configuration contradicts the assumptions.

Secondly, the gradients of w_i, $\frac{\partial J}{\partial w_{i,k}}$, converges to zero. According to Eq. (24), we have,

$$\|w_{i,\infty}\|^2 = \|w_{i,0}\|^2 + \varepsilon^2 \sum_{k=0}^{\infty} \big\|\tfrac{\partial J}{\partial w_{i,k}}\big\|^2 < \infty \tag{25}$$

from which it follows by using $\sum_k \frac{1}{k} = \infty$ that

$$\liminf_{k\to\infty} k \big\|\tfrac{\partial J}{\partial w_{i,k}}\big\|^2 = 0. \tag{26}$$

□

B. Proof of Theorems on OLS problem

B.1. Gradients and the Hessian matrix

The objective function in OLS problem (6) has an equivalent form:

$$J(a,w) = \tfrac{1}{2}(u - \tfrac{a}{\sigma}w)^T H (u - \tfrac{a}{\sigma}w) = \tfrac{1}{2}\|u\|_H^2 - \tfrac{w^T g}{\sigma}a + \tfrac{1}{2}a^2, \tag{27}$$

where $u = H^{-1}g$.

The gradients are:

$$\begin{aligned} \tfrac{\partial J}{\partial a} &= -\tfrac{1}{\sigma}(w^T H u - \tfrac{a}{\sigma}w^T H w) = -\tfrac{1}{\sigma}w^T g + a, && (28)\\ \tfrac{\partial J}{\partial w} &= -\tfrac{a}{\sigma}(Hu - \tfrac{a}{\sigma}Hw) + \tfrac{a}{\sigma^3}(w^T H u - \tfrac{a}{\sigma}w^T H w)Hw = -\tfrac{a}{\sigma}g + \tfrac{a}{\sigma^3}(w^T g)Hw. && (29)\end{aligned}$$

The Hessian matrix is

$$\begin{pmatrix} \frac{\partial^2 J}{\partial a^2} & \frac{\partial^2 J}{\partial a \partial w} \\ \frac{\partial^2 J}{\partial w \partial a} & \frac{\partial^2 J}{\partial w^2} \end{pmatrix} = \begin{pmatrix} 1 & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix} \tag{30}$$

where

$$\begin{aligned} A_{22} &= \tfrac{a}{\sigma^3}(w^T g)\Big[H + \tfrac{1}{w^T g}\big((Hw)g^T + g(Hw)^T\big) - \tfrac{3}{\sigma^2}(Hw)(Hw)^T\Big], && (31)\\ A_{21} &= -\tfrac{1}{\sigma}\big(g - \tfrac{1}{\sigma^2}(w^T g)Hw\big). && (32)\end{aligned}$$

The objective function $J(a,w)$ has saddle points, $\{(a^*, w^*) | a^* = 0, w^{*T} g = 0\}$. The Hessian matrix at those saddle points has at least one negative eigenvalue, i.e. the saddle points are strict. In fact, the eigenvalues at the saddle point (a^*, w^*) are $\Big\{\tfrac{1}{2}(1 \pm \sqrt{1 + 4\tfrac{\|g\|^2}{w^{*T}Hw^*}}), 0, ..., 0\Big\}$ which contains $d-2$ repeated zero, a positive and a negative eigenvalue.

On the other hand, the nontrivial critical points satisfies the relations,

$$a^* = \pm\sqrt{u^T H u}, w^* /\!/ u, \tag{33}$$

where the sign of a^* depends on the direction of u, w^*, i.e. $sign(a^*) = sign(u^T w^*)$. It is easy to check that the nontrivial critical points are global minimizers. The Hessian matrix at those minimizers is diag$\big(1, \tfrac{\|u\|^2}{\|w^*\|^2}H^*\big)$ where the matrix H^* is

$$H^* = H - \tfrac{Huu^T H}{u^T H u} \tag{34}$$

which is positive semi-definite and has a zero eigenvalue with eigenvector u, i.e. $H^* u = 0$. The following lemma, similar to the well-known Cauchy interlacing theorem, gives an estimate of eigenvalues of H^*.

Lemma B.1. *If H is positive definite and H^* is defined as $H^* = H - \frac{Huu^T H}{u^T H u}$, then the eigenvalues of H and H^* satisfy the following inequalities:*

$$0 = \lambda_1(H^*) < \lambda_1(H) \leq \lambda_2(H^*) \leq \lambda_2(H) \leq ... \leq \lambda_d(H^*) \leq \lambda_d(H). \tag{35}$$

Here $\lambda_i(H)$ means the i-th smallest eigenvalue of H.

Proof. (1) According to the definition, we have $H^* u = 0$, and for any $x \in \mathbb{R}^d$,

$$x^T H^* x = x^T H x - \tfrac{(x^T H u)^2}{u^T H u} \in [0, x^T H x], \tag{36}$$

which implies H^* is positive semi-definite, and $\lambda_i(H^*) \geq \lambda_1(H^*) = 0$. Furthermore, we have the following equality:

$$x^T H^* x = \min_{t \in \mathbb{R}} \|x - tu\|_H^2. \tag{37}$$

(2) We will prove $\lambda_i(H^*) \leq \lambda_i(H)$ for all i, $1 \leq i \leq d$. In fact, using the Min-Max Theorem, we have

$$\lambda_i(H^*) = \min_{dimV=i} \max_{x \in V} \frac{x^T H^* x}{\|x\|^2} \leq \min_{dimV=i} \max_{x \in V} \frac{x^T H x}{\|x\|^2} = \lambda_i(H).$$

(3) We will prove $\lambda_i(H^*) \geq \lambda_{i-1}(H)$ for all i, $2 \leq i \leq d$. In fact, using the Max-Min Theorem, we have

$$\begin{aligned}\lambda_i(H^*) &= \max_{dimV=n-i+1} \min_{x \in V} \frac{x^T H^* x}{\|x\|^2} = \max_{dimV=n-i+1, u \perp V} \min_{x \in V} \min_{t \in \mathbb{R}} \frac{\|x-tu\|_H^2}{\|x\|^2} \\ &\geq \max_{dimV=n-i+1, u \perp V} \min_{x \in V} \min_{t \in \mathbb{R}} \frac{\|x-tu\|_H^2}{\|x-tu\|^2} \\ &= \max_{dimV=n-i+1} \min_{y \in span\{V,u\}} \frac{\|y\|_H^2}{\|y\|^2}, y = x - tu \\ &\geq \max_{dimV=n-(i-1)+1} \min_{y \in V} \frac{y^T H y}{\|y\|^2} = \lambda_{i-1}(H),\end{aligned}$$

where we have used the fact that $x \perp u$, $\|x - tu\|^2 = \|x\|^2 + t^2\|u\|^2 \geq \|x\|^2$. □

There are several corollaries related to the spectral property of H^*. We first give some definitions. Since H^* is positive semi-definite, we can define the H^*-seminorm.

Definition B.2. *The H^*-seminorm of a vector x is defined as $\|x\|_{H^*} := x^T H^* x$. $\|x\|_{H^*} = 0$ if and only if x is parallel to u.*

Definition B.3. *The pseudo-condition number of H^* is defined as $\kappa^*(H^*) := \frac{\lambda_d(H^*)}{\lambda_2(H^*)}$.*

Definition B.4. *For any real number ε, the pseudo-spectral radius of the matrix $I - \varepsilon H^*$ is defined as $\rho^*(I - \varepsilon H^*) := \max_{2 \leq i \leq d} |1 - \varepsilon \lambda_i(H^*)|$.*

The following corollaries are direct consequences of Lemma B.1, hence we omit the proofs.

Corollary B.5. *The pseudo-condition number of H^* is less than or equal to the condition number of H :*

$$\kappa^*(H^*) := \frac{\lambda_d(H^*)}{\lambda_2(H^*)} \leq \frac{\lambda_d(H)}{\lambda_1(H)} =: \kappa(H), \tag{38}$$

where the equality holds if and only if $u \perp span\{v_1, v_d\}$, v_i is the eigenvector of H corresponding to the eigenvalue $\lambda_i(H)$.

Corollary B.6. *For any vector $x \in \mathbb{R}^d$ and any real number ε, we have $\|(I - \varepsilon H^*)x\|_{H^*} \leq \rho^*(I - \varepsilon H^*)\|x\|_{H^*}$.*

Corollary B.7. *For any positive number $\varepsilon > 0$, we have*

$$\rho^*(I - \varepsilon H^*) \leq \rho(I - \varepsilon H), \tag{39}$$

where the inequality is strict if $u^T v_i \neq 0$ for $i = 1, d$.

It is obvious that the inequality in Eq. (38) and Eq. (39) is strict for almost all u with respect to the Lebesgue measure. Particularly, if the spectral gap $\lambda_2(H) - \lambda_1(H)$ or $\lambda_d(H) - \lambda_{d-1}(H)$ is large, the condition number $\kappa^*(H^*)$ could be much smaller than $\kappa(H)$.

B.2. Scaling property

The dynamical system defined in Eq. (7)-(8) is completely determined by a set of configurations $\{H, u, a_0, w_0, \varepsilon_a, \varepsilon\}$. It is easy to check the system has the following scaling property:

Lemma B.8 (Scaling property)**.** *Suppose $\mu \neq 0, \gamma \neq 0, r \neq 0, Q^T Q = I$, then*

(1) The configurations $\{\mu Q^T H Q, \frac{\gamma}{\sqrt{\mu}} Q u, \gamma a_0, \gamma Q w_0, \varepsilon_a, \varepsilon\}$ and $\{H, u, a_0, w_0, \varepsilon_a, \varepsilon\}$ are equivalent.

(2) The configurations $\{H, u, a_0, w_0, \varepsilon_a, \varepsilon\}$ and $\{H, u, a_0, r w_0, \varepsilon_a, r^2 \varepsilon\}$ are equivalent.

B.3. Proof of Theorem 3.3

Recall the BNGD iterations

$$\begin{aligned} a_{k+1} &= a_k + \varepsilon_a \Big(\tfrac{w_k^T g}{\sigma_k} - a_k \Big), \\ w_{k+1} &= w_k + \varepsilon \tfrac{a_k}{\sigma_k} \Big(g - \tfrac{w_k^T g}{\sigma_k^2} H w_k \Big). \end{aligned}$$

The scaling property simplify our analysis by allowing us to set, for example, $\|u\| = 1$ and $\|w_0\| = 1$. In the rest of this section, we only set $\|u\| = 1$.

For the step size of a, it is easy to check that a_k tends to infinity with $\varepsilon_a > 2$ and initial value $a_0 = 1, w_0 = u$. Hence we only consider $0 < \varepsilon_a < 2$, which make the iteration of a_k bounded by some constant C_a.

Lemma B.9 (Boundedness of a_k)**.** *If the step size $0 < \varepsilon_a < 2$, then the sequence a_k is bounded for any $\varepsilon > 0$ and any initial value (a_0, w_0).*

Proof. Define $\alpha_k := \frac{w_k^T g}{\sigma_k}$, which is bounded by $|\alpha_k| \leq \sqrt{u^T H u} =: C$, then

$$\begin{aligned} a_{k+1} &= (1-\varepsilon_a) a_k + \varepsilon_a \alpha_k \\ &= (1-\varepsilon_a)^{k+1} a_0 + (1-\varepsilon_a)^k \varepsilon_a \alpha_0 + ... + (1-\varepsilon_a) \varepsilon_a \alpha_{k-1} + \varepsilon_a \alpha_k. \end{aligned}$$

Since $|1-\varepsilon_a| < 1$, we have $|a_{k+1}| \leq |a_0| + 2C \sum_{i=0}^{k} |1-\varepsilon_a|^i \leq |a_0| + 2C \frac{1}{1-|1-\varepsilon_a|}$. □

According to the iterations (40), we have

$$u - \tfrac{w_k^T g}{\sigma_k^2} w_{k+1} = \Big(I - \varepsilon \tfrac{a_k}{\sigma_k} \tfrac{w_k^T g}{\sigma_k^2} H \Big) \Big(u - \tfrac{w_k^T g}{\sigma_k^2} w_k \Big). \tag{40}$$

Define

$$\begin{aligned} e_k &:= u - \tfrac{w_k^T g}{\sigma_k^2} w_k, & (41) \\ q_k &:= u^T H u - \tfrac{(w_k^T g)^2}{\sigma_k^2} = \|e_k\|_H^2 \geq 0, & (42) \\ \hat{\varepsilon}_k &:= \varepsilon \tfrac{a_k}{\sigma_k} \tfrac{w_k^T g}{\sigma_k^2}, & (43) \end{aligned}$$

and using the property $\frac{w^T g}{\sigma_k^2} = \underset{t}{\operatorname{argmin}} \|u - tw\|_H$, and the property of H-norm, we have

$$q_{k+1} \leq \Big\| u - \tfrac{w_k^T g}{\sigma_k^2} w_{k+1} \Big\|_H^2 = \|(I - \hat{\varepsilon}_k H) e_k\|_H^2 \leq \rho(I - \hat{\varepsilon}_k H)^2 q_k. \tag{44}$$

Therefore we have the following lemma to make sure the iteration converge:

Lemma B.10. *Let $0 < \varepsilon_a < 2$. If there are two positive numbers ε^- and $\hat{\varepsilon}^+$, and the effective step size $\hat{\varepsilon}_k$ satisfies*

$$0 < \tfrac{\varepsilon^-}{\|w_k\|^2} \leq \hat{\varepsilon}_k \leq \hat{\varepsilon}^+ < \tfrac{2}{\lambda_{max}} \tag{45}$$

for all k large enough, then the iterations (40) converge to a minimizer.

Proof. Without loss of generality, we assume $\frac{\varepsilon^-}{\|w_k\|^2} < \frac{1}{\lambda_{max}}$ and the inequality (45) is satisfied for all $k \geq 0$. We will prove $\|w_k\|$ converges and the direction of w_k converges to the direction of u.

(1) Since $\|w_k\|$ is always increasing, we only need to prove it is bounded. We have,

$$\|w_{k+1}\|^2 = \|w_k\|^2 + \varepsilon^2 \tfrac{a_k^2}{\sigma_k^2}\|He_k\|^2 \tag{46}$$

$$= \|w_0\|^2 + \varepsilon^2 \sum_{i=0}^{k} \tfrac{a_i^2}{\sigma_i^2}\|He_i\|^2 \tag{47}$$

$$\leq \|w_0\|^2 + \varepsilon^2 \lambda_{max} \sum_{i=0}^{k} \tfrac{a_i^2}{\sigma_i^2} q_i \tag{48}$$

$$\leq \|w_0\|^2 + \varepsilon^2 \tfrac{\lambda_{max} C_a^2}{\lambda_{min}} \sum_{i=0}^{k} \tfrac{q_i}{\|w_i\|^2}. \tag{49}$$

The inequality in last lines are based on the fact that $\|He_i\|^2 \leq \lambda_{max}\|e_i\|_H^2$, and $|a_k|$ are bounded by a constant C_a. Next, we will prove $\sum_{i=0}^{\infty} \frac{q_i}{\|w_i\|^2} < \infty$, which implies $\|w_k\|$ are bounded.

According to the estimate Eq. (44), we have

$$q_{k+1} \leq \max_i \{|1 - \hat{\varepsilon}^+ \lambda_i|^2, |1 - \tfrac{\varepsilon^- \lambda_i}{\|w_k\|^2}|^2\} q_k \tag{50}$$

$$\leq \max\{1 - \gamma^+, 1 - \tfrac{\varepsilon^- \lambda_{min}}{\|w_k\|^2}\} q_k, \tag{51}$$

where $1 - \gamma^+ = \max_i\{|1 - \hat{\varepsilon}^+ \lambda_i|^2\} \in (0, 1)$. Using the definition of q_k, we have

$$q_k - q_{k+1} \geq \tfrac{\min\{\gamma^+ \|w_0\|^2, \varepsilon^- \lambda_{min}\}}{\|w_k\|^2} q_k =: \tfrac{C q_k}{\|w_k\|^2} \geq 0. \tag{52}$$

Since q_k is bounded in $[0, u^T H u]$, summing both side of the inequality, we get the bound of the infinite series $\sum_k \frac{q_k}{\|w_k\|^2} \leq \frac{u^T H u}{C} < \infty$.

(2) Since $\|w_k\|$ is bounded, we denote $\hat{\varepsilon}^- := \frac{\varepsilon^-}{\|w_\infty\|^2}$, and define $\rho := \max_i\{|1 - \hat{\varepsilon}^{\pm} \lambda_i|\} \in (0, 1)$, then the inequality (44) implies $q_{k+1} \leq \rho^2 q_k$. As a consequence, q_k tends to zero, which implies the direction of w_k converges to the direction of u.

(3) The convergence of a_k is a consequence of w_k converging.

□

Since a_k is bounded, we assume $|a_k| < \tilde{C}_a \sqrt{u^T H u}$, $\tilde{C}_a \geq 1$, and define $\varepsilon_0 := \frac{1}{2\tilde{C}_a \kappa \lambda_{max}}$. The following lemma gives the convergence for small step size.

Lemma B.11. *If the initial values (a_0, w_0) satisfies $a_0 w_0^T g > 0$, and step size satisfies $\varepsilon_a \in (0, 1], \varepsilon/\|w_0\|^2 < \varepsilon_0$, then the sequence (a_k, w_k) converges to a global minimizer.*

Remark 1: If we set $a_0 = 0$, then we have $w_1 = w_0, a_1 = \varepsilon_a \frac{w_0^T g}{\sigma_0}$, hence $a_1 w_1^T g > 0$ provided $w_0^T g \neq 0$.

Remark 2: For the case of $\varepsilon_a \in (1, 2)$, if the initial value satisfies an additional condition $0 < |a_0| \leq \varepsilon_a \frac{|w_0^T g|}{\sigma_0}$, then we have (a_k, w_k) converging to a global minimizer as well.

Proof. Without loss of generality, we only consider the case of $a_0 > 0, w_0^T g > 0, \|w_0\| \geq 1$.

(1) We will prove $a_k > 0, w_k^T g > 0$ for all k. Denote $y_k := w_k^T g$, $\delta = \frac{\|g\|}{4\kappa}$.

On the one hand, if $a_k > 0, 0 < y_k < 2\delta$, then

$$y_{k+1} \geq y_k + \varepsilon \tfrac{a_k}{\sigma_k} \tfrac{\|g\|^2}{2} \geq y_k. \tag{53}$$

On the other hand, when $a_k > 0, y_k > 0, \varepsilon < \varepsilon_0$, we have

$$y_{k+1} \ge \varepsilon \frac{a_k \|g\|^2}{\sigma_k} + y_k \left(1 - \varepsilon \frac{a_k}{\sigma_k^2} \sqrt{g^T H g}\right) \ge \frac{1}{2} y_k, \tag{54}$$

$$a_{k+1} \ge \min\{a_k, y_k / \sigma_k\}. \tag{55}$$

As a consequence, we have $a_k > 0, y_k \ge \delta_y := \min\{y_0, \delta\}$ for all k by induction.

(2) We will prove the effective step size $\hat{\varepsilon}_k$ satisfies the condition in Lemma B.10.

Since a_k is bounded, $\varepsilon < \varepsilon_0$, we have

$$\hat{\varepsilon}_k := \varepsilon \frac{a_k}{\sigma_k} \frac{w_k^T g}{\sigma_k^2} \le \frac{\varepsilon \tilde{C}_a \lambda_{max}}{\lambda_{min} \|w_k\|^2} \le \varepsilon \tilde{C}_a \kappa =: \hat{\varepsilon}^+ < \frac{1}{2\lambda_{max}}, \tag{56}$$

and

$$q_{k+1} \le (1 - \hat{\varepsilon}_k \lambda_{min})^2 q_k \le (1 - \hat{\varepsilon}_k \lambda_{min}) q_k < q_k. \tag{57}$$

which implies $\frac{w_{k+1}^T g}{\sigma_{k+1}} \ge \frac{w_k^T g}{\sigma_k} \ge \frac{w_0^T g}{\sigma_0}$. Furthermore, we have $a_k \ge \min\{a_0, \frac{w_0^T g}{\sigma_0}\}$, and there is a positive constant $\varepsilon^- > 0$ such that

$$\hat{\varepsilon}_k \ge \varepsilon \frac{a_k}{\lambda_{max} \|w_k\|^2} \frac{w_k^T g}{\sigma_k} \ge \frac{\varepsilon^-}{\|w_k\|^2}. \tag{58}$$

(3) Employing the Lemma B.10, we conclude that (a_k, w_k) converges to a global minimizer. □

Lemma B.12. *If step size satisfies $\varepsilon_a \in (0, 1], \varepsilon / \|w_0\|^2 < \varepsilon_0$, then the sequence (a_k, w_k) converges.*

Proof. Thanks to Lemma B.11, we only need to consider the case of $a_k w_k^T g \le 0$ for all k, and we will prove the iteration converges to a saddle point in this case. Since the case of $a_k = 0$ or $w_k^T g = 0$ is trivial, we assume $a_k w_k^T g < 0$ below. More specifically , we will prove $|a_{k+1}| < r|a_k|$ for some constant $r \in (0, 1)$, which implies convergence to a saddle point.

(1) If a_k and a_{k+1} have a same sign, hence different sign with $w_k^T g$, then we have $|a_{k+1}| = |1 - \varepsilon_a||a_k| - \varepsilon_a |w_k^T g| / \sigma_k \le |1 - \varepsilon_a||a_k|$.

(2) If a_k and a_{k+1} have different signs, then we have

$$\frac{|w_k^T g|}{|a_k \sigma_k|} \le \varepsilon \frac{1}{\sigma_k^2} \left(\|g\|^2 - \frac{w_k^T g}{\sigma_k^2} g^T H w_k\right) \le 2\varepsilon \kappa \lambda_{max} < 1. \tag{59}$$

Consequently, we get

$$\frac{|a_{k+1}|}{|a_k|} = \varepsilon_a \frac{|w_k^T g|}{|a_k \sigma_k|} - (1 - \varepsilon_a) \le 2\varepsilon \varepsilon_a \kappa \lambda_{max} - (1 - \varepsilon_a) < \varepsilon_a \le 1. \tag{60}$$

(3) Setting $r := \max(|1 - \varepsilon_a|, 2\varepsilon\varepsilon_a \kappa \lambda_{max} - (1 - \varepsilon_a))$, we finish the proof. □

To simplify our proofs for Theorem 3.3, we give two lemmas which are obvious but useful.

Lemma B.13. *If positive series f_k, h_k satisfy $f_{k+1} \le r f_k + h_k, r \in (0, 1)$ and $\lim_{k \to \infty} h_k = 0$, then $\lim_{k \to \infty} f_k = 0$.*

Proof. It is obvious, because the series b_k defined by $b_{k+1} = r b_k + h_k, b_0 > 0$, tends to zeros. □

Lemma B.14 (Separation property)**.** *For δ_0 small enough, the set $S := \{w | y^2 q < \delta_0, \|w\| \ge 1\}$ is composed by two separated parts: S_1 and S_2, $dist(S_1, S_2) > 0$, where in the set S_1 one has $y^2 < \delta_1, q > \delta_2$, and in S_2 one has $q < \delta_2, y^2 > \delta_1$ for some $\delta_1 > 0, \delta_2 > 0$. Here $y := w^T g, q := u^T H u - \frac{(w^T H u)^2}{w^T H w} = u^T H u - \frac{y^2}{w^T H w}$.*

Proof. The proof is based on H being positive. The geometric meaning is illustrated in Figure 5. □

Corollary B.15. *If $\lim_{k \to \infty} \|w_{k+1} - w_k\| = 0$, and $\lim_{k \to \infty} (w_k^T g)^2 q_k = 0$, then either $\lim_{k \to \infty} (w_k^T g)^2 = 0$ or $\lim_{k \to \infty} q_k = 0$.*

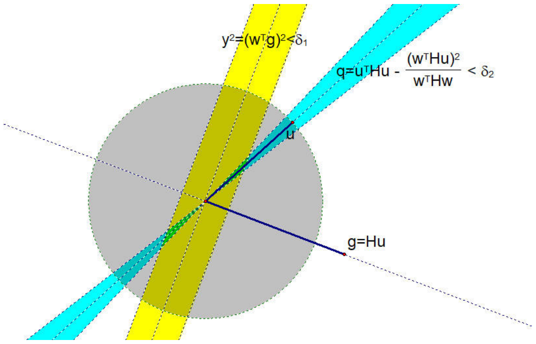

Figure 5. The geometric meaning of the separation property

Proof. Denote $y_k := w_k^T g$. According to the separation property (Lemma B.14), we can chose a $\delta_0 > 0$ small enough such that the separated parts of the set $S := \{w | y^2 q < \delta_0, \|w\| \geq 1\}$, S_1 and S_2, have $dist(S_1, S_2) > 0$.

Because $y_k^2 q_k$ tends to zero, we have w_k belongs to S for k large enough, for instance $k > k_1$. On the other hand, because $\|w_{k+1} - w_k\|$ tends to zero, we have $\|w_{k+1} - w_k\| < dist(S_1, S_2)$ for k large enough, for instance $k > k_2$. Then consider $k > k_3 := \max(k_1, k_2)$, we have all w_k belongs to the same part S_1 or S_2.

If $w_k \in S_1$, ($q_k > \delta_2$), for all $k > k_3$, then we have $\lim\limits_{k\to\infty} (w_k^T g)^2 = 0$.

On the other hand, if $w_k \in S_2$, ($y_k^2 > \delta_1$), for all $k > k_3$, then we have $\lim\limits_{k\to\infty} q_k = 0$.

□

Theorem B.16. *Let $\varepsilon_a \in (0, 1]$ and $\varepsilon > 0$. The sequence (a_k, w_k) converges for any initial value (a_0, w_0).*

Proof. We will prove $\|w_k\|$ converges, and then prove (a_k, w_k) converges as well.

(1) We prove that $\|w_k\|$ is bounded and hence converges.

In fact, according to the Lemma B.12, once $\|w_k\|^2 \geq \varepsilon/\varepsilon_0$ for some k, the rest of the iteration will converge, hence $\|w_k\|$ is bounded.

(2) We prove $\lim\limits_{k\to\infty} \|w_{k+1} - w_k\| = 0$, and $\lim\limits_{k\to\infty} (w_k^T g)^2 q_k = 0$.

The convergence of $\|w_k\|$ implies $\sum_k a_k^2 q_k$ is summable. As a consequence,

$$\lim_{k\to\infty} a_k^2 p_k = 0, \lim_{k\to\infty} a_k e_k = 0, \tag{61}$$

and $\lim\limits_{k\to\infty} \|w_{k+1} - w_k\|$ = 0. In fact, we have

$$\|w_{k+1} - w_k\|^2 = \varepsilon^2 \tfrac{a_k^2}{\sigma^2} \|He_k\|^2 \leq \tfrac{\lambda_{max}\varepsilon^2}{\lambda_{min}^2} a_k^2 q_k \to 0. \tag{62}$$

Consider the iteration of series $|a_k - w_k^T g/\sigma_k|$,

$$\begin{aligned} \left|a_{k+1} - \tfrac{w_{k+1}^T g}{\sigma_{k+1}}\right| &\leq \left|a_{k+1} - \tfrac{w_{k+1}^T g}{\sigma_k}\right| + \left|\tfrac{w_{k+1}^T g}{\sigma_k} - \tfrac{w_{k+1}^T g}{\sigma_{k+1}}\right| \\ &\leq (1 - \varepsilon_a)\left|a_k - \tfrac{w_k^T g}{\sigma_k}\right| + \varepsilon \tfrac{|a_k g^T H e_k|}{\sigma_k^2} + \tfrac{|w_{k+1}^T g|}{(\sigma_k \sigma_{k+1})} |\sigma_{k+1} - \sigma_k| \\ &\leq (1 - \varepsilon_a)\left|a_k - \tfrac{w_k^T g}{\sigma_k}\right| + \varepsilon \tfrac{\|g\|_H \|a_k e_k\|_H}{\sigma_k^2} + \tfrac{|w_{k+1}^T g|}{(\sigma_k \sigma_{k+1})} \varepsilon \tfrac{\lambda_{max}}{\sigma_k} \|a_k e_k\|_H \\ &\leq (1 - \varepsilon_a)\left|a_k - \tfrac{w_k^T g}{\sigma_k}\right| + 2C\|a_k e_k\|_H. \end{aligned} \tag{63}$$

The constant C in Eq. (63) can be chosen as $C = \frac{\varepsilon\lambda_{max}\|u\|_H}{\lambda_{min}\|w_0\|^2}$. Since $\|a_k e_k\|_H$ tends to zero, we can use Lemma B.13 to get $\lim\limits_{k\to\infty} |a_k - w_k^T g/\sigma_k| = 0$. Combine the equation (61), then we have $\lim\limits_{k\to\infty}(w_k^T g)^2 p_k = 0$.

(3) According to the Corollary B.15, we have either $\lim\limits_{k\to\infty} y_k^2 = 0$, or $\lim\limits_{k\to\infty} q_k = 0$. In the former case, the iteration of (a_k, w_k) converges to a saddle point. However, in the latter case, (a_k, w_k) converges to a global minimizer. In both cases we have (a_k, w_k) converges.

□

To finish the proof of Theorem 3.3, we have to demonstrate the special case of $\varepsilon_a = 1$ where the set of initial values such that BN iteration converges to saddle points is of Lebeguse measure zero. We leave this demonstration in next section where we consider the case of $\varepsilon_a \geq 1$.

B.4. Impossibility of converging to strict saddle points

In this section, we will prove the set of initial values such that BN iteration converges to saddle points is of Lebesgue measure zero, as long as $\varepsilon_a \geq 1$. The tools in our proof is similar to the analysis of gradient descent on non-convex objectives (Lee et al., 2016; Panageas & Piliouras, 2017). In addition, we used the real analytic property of the BN loss function (27).

For brevity, here we denote $x := (a, w)$ and let $\varepsilon_a = \varepsilon$, then the BN iteration can be rewritten as

$$x_{n+1} = T(x_n) := x_n - \varepsilon\nabla J(x_n).$$

Lemma B.17. *If $A \subset T(\mathbb{R}^d/\{0\})$ is a measure zero set, then the preimage $T^{-1}(A)$ is of measure zero as well.*

Proof. Since T is smooth enough, according to Theorem 3 of Ponomarev (1987), we only need to prove the Jacobian of $T(x)$ is nonzero for almost all $x \in \mathbb{R}^d$. In other words, the set $\{x : \det(I - \varepsilon\nabla^2 J(x)) = 0\}$ is of measure zero. This is true because the function $\det(I - \varepsilon\nabla^2 J(x))$ is a real analytic function of $x \in \mathbb{R}^d/\{0\}$. (Details of properties of real analytic functions can be found in Krantz & Parks (2002)).

□

Lemma B.18. *Let $f : X \to \mathbb{R}$ be twice continuously differentiable in an open set $X \subset \mathbb{R}^d$ and $x^* \in X$ be a stationary point of f. If $\varepsilon > 0$, $\det(I - \varepsilon\nabla^2 f(x^*)) \neq 0$ and the matrix $\nabla^2 f(x^*)$ has at least a negative eigenvalue, then there exist a neighborhood U of x^* such that the following set B has measure zero,*

$$B := \{x_0 \in U : x_{n+1} = x_n - \varepsilon\nabla f(x_n) \in U, \forall n \geq 0\}. \tag{64}$$

Proof. The detailed proof is similar to Lee et al. (2016); Panageas & Piliouras (2017).

Define the transform function as $F(x) := x - \varepsilon\nabla f(x)$. Since $\det(I - \varepsilon\nabla^2 f(x^*)) \neq 0$, according to the inverse function theorem, there exist a neighborhood U of x^* such that T has differentiable inverse. Hence T is a local C^1 diffeomorphism, which allow us to use the central-stable manifold theorem (Shub, 2013). The negative eigenvalues of $\nabla^2 f(x^*)$ indicates $\lambda_{max}(I - \varepsilon\nabla^2 f(x^*)) > 1$ and the dimension of the unstable manifold is at least one, which implies the set B is on a lower dimension manifold hence B is of measure zero.

□

Lemma B.19. *If $\varepsilon_a = \varepsilon \geq 1$, then the set of initial values such that BN iteration converges to saddle points is of Lebeguse measure zero.*

Proof. We will prove this argument using Lemma B.17 and Lemma B.18. Denote the saddle points set as $W := \{(a^*, w^*) : a^* = 0, w^{*T} g = 0\}$. The basic point is that the saddle point $x^* := (a^*, w^*)$ of the BN loss function (27) has eigenvalues $\left\{\frac{1}{2}(1 \pm \sqrt{1 + 4\frac{\|g\|^2}{w^{*T} H w^*}}), 0, ..., 0\right\}$ of the Hessian matrix.

(1) For each saddle point $x^* := (a^*, w^*)$ of BN loss function, $\varepsilon \geq 1$ is enough to allow us to use Lemma B.18. Hence there exist a neighborhood U_{x^*} of x^* such that the following set B_{x^*} is of measure zero,

$$B_{x^*} := \{x_0 \in U_{x^*} : x_n \in U_{x^*}, \forall n \geq 1\}. \tag{65}$$

(2) The neighborhoods U_{x^*} of all $x^* \in W$ forms a cover of W, hence, according to Lindelöf's open cover lemma, there are countable neighborhoods $\{U_i : i = 1, 2, ...\}$ cover W, i.e. $U := \cup_i U_i \supseteq W$. As a consequence, the following set A_0 is of measure zero,

$$A_0 := \cup_i B_i = \cup_i \{x_0 \in U_i : x_n \in U_i, \forall n \geq 1\}. \tag{66}$$

(3) Define $A_{m+1} := T^{-1}(A_m) = \{x \in \mathbb{R}^d : T(x) \in A_m\}, m \geq 0$. According to Lemma B.17, we have all A_m and $\cup_m A_m$ are of measure zero.

(4) Since each initial value x_0 such that the iteration converges to a saddle point must be contained in some set A_m, we finish the proof.

□

Combine the results of Lemma B.19, scaling property 3.2 and the convergence theorem B.16, we have the following theorem directly.

Theorem B.20. *If $\varepsilon_a = 1, \varepsilon \geq 0$, then the BN iteration (7)-(8) converges to global minimizers for almost all initial values.*

B.5. Convergence rate

In section B.3, we encountered the following estimate for $e_k = u - \frac{w_k^T g}{\sigma_k^2} w_k$

$$\|e_{k+1}\|_H \leq \rho(I - \hat{\varepsilon}_k H)\|e_k\|_H. \tag{67}$$

We can improve the convergence rate of the above if H^* has better spectral property. This is the content of Theorem 3.4 and the following lemma proves this.

Lemma B.21. *The following inequality holds,*

$$(1 - \delta_k)\|e_{k+1}\|_H \leq \Big(\rho^*(I - \hat{\varepsilon}_k H^*) + \delta_k\Big)\|e_k\|_H, \tag{68}$$

where $\delta_k := \frac{\lambda_{max}\varepsilon|a_k|}{\sigma_k^2}\|e_k\|_H$.

Proof. The case of $w_k^T g = 0$ is trivial, hence we assume $w_k^T g \neq 0$ in the following proof. Rewrite the iteration on w_k as the following equality,

$$u - \frac{w_k^T g}{\sigma_k^2} w_{k+1} = (I - \hat{\varepsilon}_k H)e_k = (I - \hat{\varepsilon}_k H^*)e_k - \hat{\varepsilon}_k\Big(1 - \frac{(w_k^T g)^2}{u^T H u \sigma_k^2}\Big)Hu. \tag{69}$$

Then we will use the properties of H^*-seminorm to prove our argument.

(1) Estimate the H^*-seminorm on the right hand of Eq. (69).

$$\begin{aligned}
\|\text{right}\|_{H^*} &\leq \|(I - \hat{\varepsilon}_k H^*)e_k\|_{H^*} + |\hat{\varepsilon}_k|\Big(1 - \frac{(w_k^T g)^2}{u^T H u \sigma_k^2}\Big)\|Hu\|_{H^*} && (70)\\
&\leq \rho^*(I - \hat{\varepsilon}_k H^*)\|e_k\|_{H^*} + \frac{\lambda_{max}|\hat{\varepsilon}_k|}{\sqrt{u^T H u}}\|e_k\|_H^2 && (71)\\
&= \rho^*(I - \hat{\varepsilon}_k H^*)\frac{|w_k^T g|}{\sqrt{u^T H u}\sigma_k}\|e_k\|_H + \frac{\lambda_{max}\varepsilon|a_k w_k^T g|}{\sqrt{u^T H u}\sigma_k^3}\|e_k\|_H^2 && (72)\\
&= \frac{|w_k^T g|}{\sqrt{u^T H u}\sigma_k}\Big(\rho^*(I - \hat{\varepsilon}_k H^*) + \delta_k\Big)\|e_k\|_H. && (73)
\end{aligned}$$

(2) Estimate the H^*-seminorm on the left hand of equation (69). Using the H-norm on the iteration of w_k, we have

$$\sigma_{k+1} = \|w_k + \varepsilon\frac{a_k}{\sigma_k}He_k\|_H \geq \sigma_k - \varepsilon\frac{\lambda_{max}|a_k|}{\sigma_k}\|e_k\|_H. \tag{74}$$

Consequently, we have

$$\|\text{left}\|_{H^*} = \frac{|w_k^T g|}{\sqrt{u^T H u}\sigma_k}\frac{\sigma_{k+1}}{\sigma_k}\|e_{k+1}\|_H \geq \frac{|w_k^T g|}{\sqrt{u^T H u}\sigma_k}(1 - \delta_k)\|e_{k+1}\|_H. \tag{75}$$

(3) Combining (1) and (2), we finish the proof. □

Then we give the proof of Theorem 3.4.

Proof of Theorem 3.4. Firstly, the Lemma B.21 implies the second part of Theorem 3.4 which is the special case of $\delta_k < \delta < 1$.

Secondly, if $\hat{\varepsilon} < \varepsilon^*_{max}$, then $\rho^*(I - \hat{\varepsilon} H^*) < 1$. Since (a_k, w_k) converges to a minimizer, δ_k must converge to zero and the coefficient $\frac{\rho^*(I - \hat{\varepsilon}_k H^*) + \delta_k}{(1-\delta_k)}$ must less than a number $\hat{\rho} \in (0,1)$ when k is large enough which results in the linear convergence of $\|e_k\|_H$. □

Now, we turn to the convergence of the loss function which can be rewritten as $J_k = \frac{1}{2}\|\tilde{e}_k\|^2_H$ with $\tilde{e}_k = u - \frac{a_k}{\sigma_k} w_k$. There is an useful equality between $\|\tilde{e}_k\|^2_H$ and $\|e_k\|^2_H$:

$$\|\tilde{e}_k\|^2_H = \|e_k\|^2_H + \left(a_k - \frac{w_k^T g}{\sigma_k}\right)^2. \tag{76}$$

Recalling the inequality (63) and the boundedness of a_k, we have a constant C_0 such that

$$\left|a_{k+1} - \frac{w_{k+1}^T g}{\sigma_{k+1}}\right| \leq |1 - \varepsilon_a| \left|a_k - \frac{w_k^T g}{\sigma_k}\right| + C_0 \|e_k\|_H, \tag{77}$$

which indicates that we can use the convergence of e_k to estimate the convergence of the loss value J_k. In fact we have the following lemma.

Lemma B.22. *If $\|e_k\|_H \leq C\rho^k$ for some constant C and $\rho \in (0,1)$, $\varepsilon_a \in (0,1]$, then we have*

$$\|\tilde{e}_k\|^2_H \leq C^2 \rho^{2k} + \left(C_1(1-\varepsilon_a)^k + C_2 k \gamma^k\right)^2, \tag{78}$$

where $\gamma = \max(\rho, 1 - \varepsilon_a)$, $C_1 = |a_0 - w_0^T g/\sigma_0|$ and $C_2 = CC_0$.

Proof. According to the inequality (77), we have

$$\left|a_k - \frac{w_k^T g}{\sigma_k}\right| \leq C_1(1-\varepsilon_a)^k + C_2 \sum_{i=0}^{k-1} (1-\varepsilon_a)^i \rho^{k-i} \leq C_1(1-\varepsilon_a)^k + C_2 k \gamma^k. \tag{79}$$

Put it in the Eq. (76), then we finish the proof.

□

B.6. Estimating the effective step size

Firstly, we consider the limit of effective step size $\hat{\varepsilon}$. When the iteration converges to a minimizer (a^*, w^*), the value of $\hat{\varepsilon}$ is $\hat{\varepsilon} = \frac{\varepsilon}{\|w^*\|^2}$. Without loss generality, we assume that w_k always has different direction with u during the whole course of the iterations. In fact, if w_k has the same direction with u for some k, then the iteration of w_k is trivial, i.e. $w_k = w_{k+1} = w_{k+2} = ...$, and the effective step size can be any positive number. However, this case is rare. More precisely, we have the following lemma:

Lemma B.23. *The set of initial values (a_0, w_0) such that (a_k, w_k) converges to a minimizer (a^*, w^*) with effective learning rate $\hat{\varepsilon} := \lim_{k\to\infty} \hat{\varepsilon}_k > \varepsilon^*_{max}$ and $\det(I - \hat{\varepsilon} H^*) \neq 0$ is of measure zero.*

Proof. The proof is similar to the proof of Lemma B.20. The key point is that the matrix $I - \hat{\varepsilon} H^*$ at this minimizer is non-degenerate and has an eigenvalue with its absolute value large than 1, hence there is a local unstable manifold with dimension greater than one. □

Now we consider the effective learning rate $\hat{\varepsilon}_k$ and give the proof of Proposition 3.5.

According to Lemma B.11, the effective step size $\hat{\varepsilon}_k$ has same order with $\frac{\varepsilon}{\|w_k\|^2}$ provided $a_0 w_0^T g > 0, \varepsilon/||w_0|| < \varepsilon_0$. In fact, we have

$$\frac{C_1 \varepsilon}{\|w_k\|^2} := \frac{a_0 w_0^T g}{\sigma_0} \frac{\varepsilon}{\lambda_{max}\|w_k\|^2} \leq \hat{\varepsilon}_k \leq \sqrt{u^T H u} \frac{C_a \varepsilon}{\lambda_{min}\|w_k\|^2} =: \frac{C_2 \varepsilon}{\|w_k\|^2}. \tag{80}$$

Hence, to prove the Proposition 3.5, we only need to estimate the norm of w_k.

Proof of Proposition 3.5. According to the BNGD iteration, we have (see the proof of Lemma B.10)

$$\|w_{k+1}\|^2 \leq \|w_0\|^2 + \varepsilon^2 \lambda_{max} \sum_{i=0}^{k} \tfrac{a_i^2}{\sigma_i^2} q_i. \tag{81}$$

(1) When $\frac{\varepsilon}{\|w_0\|^2} < \varepsilon_0$ (ε_0 is defined in Lemma B.11), the sequence q_k satisfies $q_{k+1} \leq (1 - \hat{\varepsilon}_k \lambda_{min}) q_k$. Hence the norm of w_k is bounded by

$$\|w_k\|^2 \leq \|w_0\|^2 + \varepsilon \kappa C_a \tfrac{\sigma_0}{w_0^T g} \sum_{i=0}^{\infty} (q_i - q_{i+1}) \leq \|w_0\|^2 + C\varepsilon, \tag{82}$$

for some constant C. As a consequence,

$$\tilde{C}_1 \varepsilon := \tfrac{C_1 \varepsilon}{\|w_0\|^2 (1 + C\varepsilon_0)} \leq \hat{\varepsilon}_k \leq \tfrac{C_2 \varepsilon}{\|w_0\|^2} =: \tilde{C}_2 \varepsilon. \tag{83}$$

(2) When ε is large enough, the increment of the norm $\|w_k\|$ at the first step is large as well. In fact, we have

$$\|w_1\|^2 - \|w_0\|^2 = \varepsilon^2 \tfrac{a_0^2}{\sigma_0^2} \|He_0\|^2 = C_3 \varepsilon^2. \tag{84}$$

Since $||g||^2 \geq \frac{w_0^T g}{\sigma_0^2} g^T H w_0$, we have $a_1 w_1^T g > a_1 w_0^T g > 0$. Choose ε to be larger than some value ε_1 such that $\frac{\varepsilon}{\|w_1\|^2} < \varepsilon_0$, then we can use the argument in (1) on (a_1, w_1). More precisely, there are two constants, C_1, C_2, such that

$$\tfrac{C_1 \varepsilon}{\|w_1\|^2} \leq \hat{\varepsilon}_k \leq \tfrac{C_2 \varepsilon}{\|w_1\|^2}. \tag{85}$$

Plugging the equation (84) into it, we have

$$\tfrac{C_1 \varepsilon_1^2}{\|w_0\|^2 + C_3 \varepsilon_1^2} \leq \tfrac{C_1 \varepsilon^2}{\|w_0\|^2 + C_3 \varepsilon^2} \leq \hat{\varepsilon}_k \varepsilon \leq \tfrac{C_2 \varepsilon^2}{\|w_0\|^2 + C_3 \varepsilon^2} \leq \tfrac{C_2}{C_3}. \tag{86}$$

□

B.7. Quantification of the insensitive interval

In this section, we estimate the magnitude of insensitive interval of step size.

The BNGD iteration with configuration $\varepsilon_a = 1, a_0 = \frac{w_0^T g}{\sigma_0}, \|w_0\| = \|u\| = 1$ implies the following equality of $\|w_k\|^2$,

$$\begin{aligned} \|w_{k+1}\|^2 &= \|w_k\|^2 + \tfrac{\varepsilon^2}{\|w_k\|^2} \tfrac{a_k^2 \|w_k\|^2}{\sigma_k^2} \left\|e_k\right\|_{H^2}^2 \\ &=: \|w_k\|^2 + \tfrac{\varepsilon^2}{\|w_k\|^2} \beta_k, \end{aligned} \tag{87}$$

where β_k is defined as $\beta_k := \frac{a_k^2 \|w_k\|^2}{\sigma_k^2} \left\|e_k\right\|_{H^2}^2$. The linear convergence results allow us to assume that β_k converges linearly to zero, i.e. $\beta_k = \beta_0 \rho^k, k \geq 0$ where $\rho \in (0,1)$ depends on ε and is self-consistently determined by the limiting effective step size, i.e. $\rho = \rho(I - \frac{\varepsilon}{\|w_\infty\|^2} H)$ is the spectral radius of $I - \frac{\varepsilon}{\|w_\infty\|^2} H$. Observed that the iteration in Eq. (87) can be regarded as a numerical scheme for solving the following ODE:

$$\xi(0) = \|w_1\|^2, \dot{\xi}(t) = \frac{\varepsilon^2 \beta_0 \rho^{2t}}{\xi(t)}, \tag{88}$$

which has solution $\xi^2(t) = \xi^2(0) + \frac{\varepsilon^2 \beta_0}{|\ln \rho|}(1 - \rho^{2t})$, the value of $\|w_k\|^2$ can be approximated by $\xi(k+1)$. Particularly, we have an approximation for $\|w_\infty\|^2$:

$$\|w_\infty\|^2 \approx \xi(\infty) = \sqrt{(1 + \varepsilon^2 \beta_0)^2 + \tfrac{\varepsilon^2 \beta_0}{|\ln \rho|}}. \tag{89}$$

To determine the value of ρ, we let ρ and ε satisfy the following relation:

$$\rho = \rho(I - \tfrac{\varepsilon}{\xi(\infty)}H) := \max_i\{|1 - \tfrac{\varepsilon}{\xi(\infty)}\lambda_i(H)|\}, \tag{90}$$

which closed the calculation of $\xi(\infty)$.

Next, we consider two limiting case: $\varepsilon \ll 1$ and $\varepsilon \gg 1$. In both case, the effective step size $\hat{\varepsilon}$ is small, and the value of ρ is related to $\rho = 1 - \frac{\varepsilon\lambda_{min}}{\xi(\infty)}$. Combine the definition of $\xi(\infty)$, then we have

$$\tfrac{\varepsilon^2\lambda_{min}^2}{(1-\rho)^2} = \xi(\infty)^2 = (1+\varepsilon^2\beta_0)^2 + \tfrac{\varepsilon^2\beta_0}{|\ln\rho|} \approx (1+\varepsilon^2\beta_0)^2 + \tfrac{\varepsilon^2\beta_0}{1-\rho}, \tag{91}$$

where the estimate of $|\ln\rho| \approx 1-\rho$, is used since ρ is closed to 1 for $\hat{\varepsilon}$ is small enough. Consequently, we have:

(1) When $\varepsilon \ll 1$, we have $\alpha^* \approx 1$, $\rho \approx 1 - \varepsilon\lambda_{min}$ and $\hat{\varepsilon} \approx \varepsilon$.

(2) When $\varepsilon \gg 1$, we have

$$\hat{\varepsilon} \approx \frac{1-\rho}{\lambda_{min}} \approx \frac{\sqrt{1+4\varepsilon^2\lambda_{min}^2}-1}{2\varepsilon^2\beta_0\lambda_{min}} = \frac{1}{\beta_0}\frac{2\lambda_{min}}{\sqrt{1+4\varepsilon^2\lambda_{min}^2}+1} \sim \tfrac{1}{\beta_0\varepsilon}. \tag{92}$$

Those results indicate the magnitude of insensitive interval of step size is proportion to the constant $\frac{1}{\beta}_0$.

Finally, we estimate the average of β_0 over w_0 and u for given H. The average value of β_0 from BNGD is defined as the following geometric average over w_0 and u, which we take to be independent and uniformly on the unit sphere $\mathbb{S}^{d-1}$,

$$\bar{\beta}_H := \mathbb{E}^G_{w_0,u}[\beta_0] := \exp\big(\mathbb{E}_{w_0,u}\ln\big[\big(\tfrac{w_0^T Hu}{w_0^T Hw_0}\big)^2\|e_0\|^2_{H^2}\big]\big). \tag{93}$$

Correspondingly, the magnitude of insensitive interval of step size is defined as Ω,

$$\Omega = \Omega_H := \mathbb{E}^G_{w_0,u}[\tfrac{\lambda^2_{max}(H)}{4\beta_0}] = \tfrac{\lambda^2_{max}(H)}{4\bar{\beta}_H}. \tag{94}$$

The numerical tests find that Ω_H highly depends on the dimension d provided the eigenvalues of H is sampled from typical distributions such as the uniform distribution on $[\lambda_{min}, \lambda_{max}]$ with $0 < \lambda_{min} < \lambda_{max}$. In fact we have the following estimations for $\bar{\beta}_H$ which implies $\bar{\beta}_H \le O(1/d)$ and $\Omega_H \ge O(d)$.

Lemma B.24. *For positive definite matrix H with minimal and maximal eigenvalues, λ_{min} and λ_{max} respectively, the $\bar{\beta}_H$ defined in (93) satisfies,*

$$\bar{\beta}_H \le \tfrac{1}{d}\tfrac{Tr[H^2]}{d}\tfrac{\lambda_{max}Tr[H]}{d}\tfrac{1}{\lambda^2_{min}}\exp\big(-\tfrac{2\ln\kappa}{\kappa-1}\big(\tfrac{Tr[H]}{d\lambda_{min}}-1\big)\big), \tag{95}$$

where $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ is the condition number of H.

Proof. The definition of $\mathbb{E}^G$ allows us to estimate each term in β separately.

(1). The inequality of arithmetic and geometric means implies $\mathbb{E}^G[(w_0^T Hu)^2] \le \mathbb{E}[(w_0^T Hu)^2] = \frac{Tr[H^2]}{d^2}$.

(2). Using the definition of $e_0 = u - \frac{w_0^T Hu}{w_0^T Hw_0}w_0$, we have

$$\begin{aligned}\mathbb{E}^G\big[\|e_0\|^2_{H^2}\big] \le \lambda_{max}\mathbb{E}\big[\|e_0\|^2_H\big] &= \lambda_{max}\mathbb{E}\big[\|u - \tfrac{w_0^T Hu}{w_0^T Hw_0}w_0\|^2_H\big] \\ &= \lambda_{max}\mathbb{E}\big[u^T Hu - \big(\tfrac{w_0^T Hu}{w_0^T Hw_0}\big)^2\big] \\ &\le \lambda_{max}\mathbb{E}\big[u^T Hu\big] = \tfrac{\lambda_{max}Tr[H]}{d}.\end{aligned}$$

(3). Since $w_0^T H w_0 \in [\lambda_{min,\lambda_{max}}]$, using the fact that $\ln(1+x) \geq \frac{\ln\kappa}{\kappa-1}, \forall x \in [0, \kappa-1]$, we have

$$\begin{aligned}\mathbb{E}^G\left[w_0^T H w_0\right] &= \exp\left(\mathbb{E}\ln(w_0^T H w_0)\right) \\ &\geq \lambda_{min}\exp\left(\mathbb{E}\tfrac{\ln\kappa}{\kappa-1}(w_0^T H w_0/\lambda_{min}-1)\right) \\ &= \lambda_{min}\exp\left(\tfrac{\ln\kappa}{\kappa-1}(\tfrac{Tr[H]}{d\lambda_{min}}-1)\right).\end{aligned} \tag{96}$$

Combine the inequities above, then we finish the proof. □

If the eigenvalues of H is sampled from a given distribution on $[\lambda_{min}, \lambda_{max}]$, the values $\frac{Tr[H]}{d}, \frac{Tr[H^2]}{d}$ are related to the distribution and not sensitive to dimension d (for d large enough), then the estimate in Lemma B.24 indicates that $\bar{\beta}_H \leq O(1/d)$ and $\Omega_H \geq O(d)$. As an example, we consider the H with eigenvalues forming an arithmetic sequence below.

Corollary B.25. *If the eigenvalues of H are $\lambda_i = \lambda_{min} + (i-1)\frac{\lambda_{max}-\lambda_{min}}{d-1}, d \geq 2$, then we have*

$$\bar{\beta}_H \leq \tfrac{(\kappa+1)^3}{\kappa^2}\tfrac{\lambda_{max}^2}{4d},\ \Omega_H \geq \frac{\kappa^2}{(\kappa+1)^3}d. \tag{97}$$

Proof. It is enough to show that $\frac{Tr[H]}{d} = \frac{(\kappa+1)\lambda_{min}}{2}, \frac{Tr[H^2]}{d^2} \leq \frac{(\kappa+1)^2\lambda_{min}^2}{2d}$. □

The Corollary B.25 indicates that larger dimensions lead to larger insensitive intervals of step size. It is interesting to note that although the lower bound of Ω_H is also related to the condition number κ, the numerical tests in section B.7.1 find the width is not sensitive to κ. In fact, one could get better lower bounds for Ω_H by better estimates on $\mathbb{E}^G(w^T H w)$. However, here we focus on the effect of dimension.

B.7.1. Numerical tests

In this section, we give some numerical tests on the BNGD iteration with $\varepsilon_a = 1, a_0 = 0$ and choices of the matrix H. The scaling property allows us to set H diagonal and the initial value w_0 having the same norm with u, $\|w_0\| = \|u\| = 1$.

Firstly, we show the difference of geometric mean(G-mean) and arithmetic mean(A-mean) in quantifying the performance of BNGD. Figure 6 gives an example of a 100-dimensional H with condition number $\kappa = 853$. The GD and MBNGD iteration are executed $k = 5000$ times where u and w_0 are randomly chosen from the unit sphere. The values of effective step size, loss $\|e_k\|_H^2$ and error $\|e_k\|$ are plotted. Furthermore, the mean values over 500 random tests are given. The results show that the G-mean converges quickly when the number of tests increase, however the A-mean does not converge as quickly and A-mean is dominated by the largest sample values. Hence we use the geometric mean in later tests.

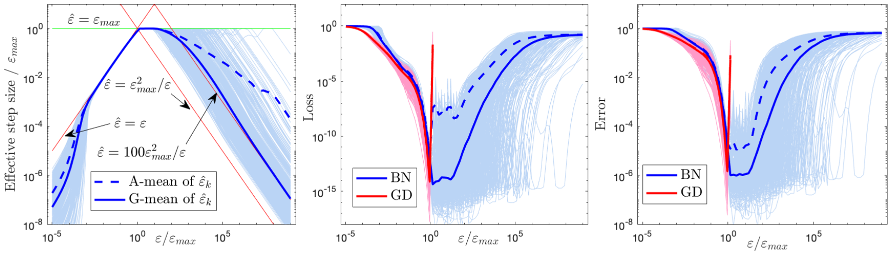

Figure 6. Test BNGD on OLS model with step size $\varepsilon_a = 1, a_0 = 0$. Parameters: H is a diagonal matrix with condition number $\kappa = 853$ (the first random test in Figure 8), u and w_0 is randomly chosen uniformly from the unit sphere in $\mathbb{R}^{100}$. The BNGD iterations are executed for $k = 5000$ steps. The bold curves are averaged over the 500 independent runs (the shadow curves).

Secondly, we test the effect of dimension d.

Figure 7 gives three typical setting of H: (a) with arithmetic progression eigenvalues, (b) with geometric progression eigenvalues and (c) with only one large eigenvalue perturbed from identity matrix. In the first two cases, the effect of dimension is observed, the large dimensions lead to large magnitude Ω of optimal step size, and the magnitude is almost proportion to the dimension d which confirm the analysis in Lemma B.24 and Corollary B.25. In the last case, the large dimensions lead to small Ω which is due to $Tr[H]/d$ and $Tr[H^2]/d$ are highly influenced by d. However, the condition number of H^* be much smaller than $\kappa(H)$, in which case leads to marked acceleration over GD.

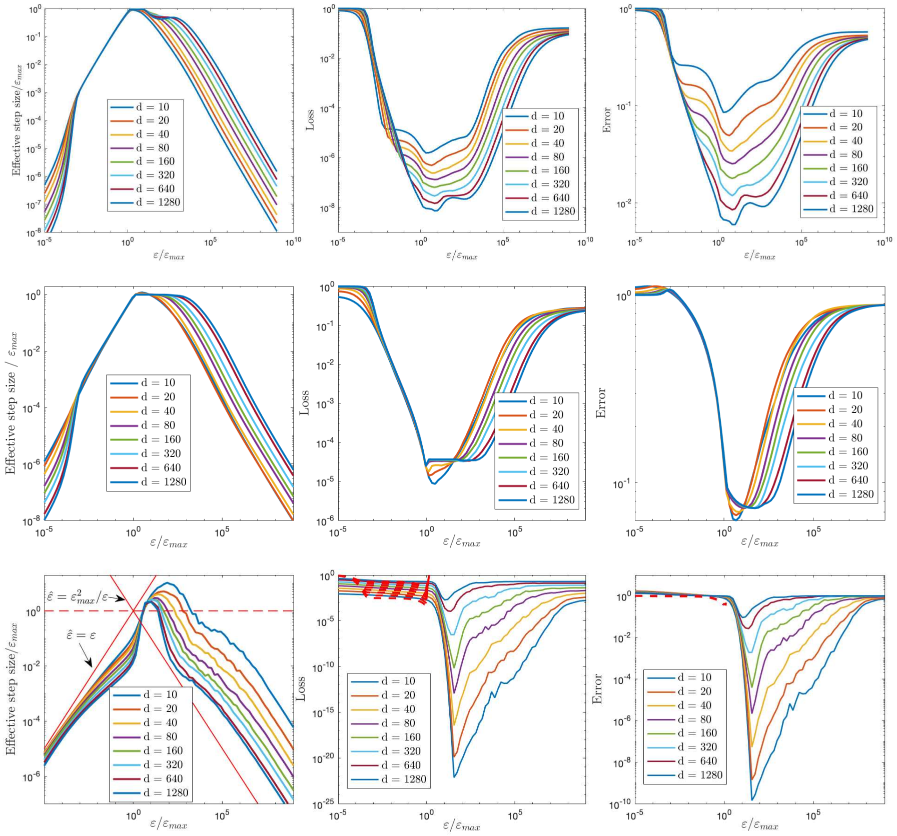

Figure 7. Tests of BNGD on OLS model with step size $\varepsilon_a = 1, a_0 = 0$. Parameters: (a, top) H = diag(linspace(1,10000,d)), (b, middle) H = diag(logspace(0,4,d)), (c, bottom) H = diag([ones(1,d-1),10000]]). u and w_0 is randomly chosen uniformly from the unit sphere in $\mathbb{R}^d$. The BNGD iterations are executed for $k = 5000$ steps. The curves are averaged over the 500 independent runs.

Finally, we test the effect of eigenvalue distributions. Figure 8 gives examples of H with different condition number but same dimension $d = 100$. When the eigenvalues are arithmetic sequences, the width of optimal learning rate is almost same over different condition numbers while the loss and error still depend on the condition number. Randomly choosing eigenvalues also exhibits this phenomenon.

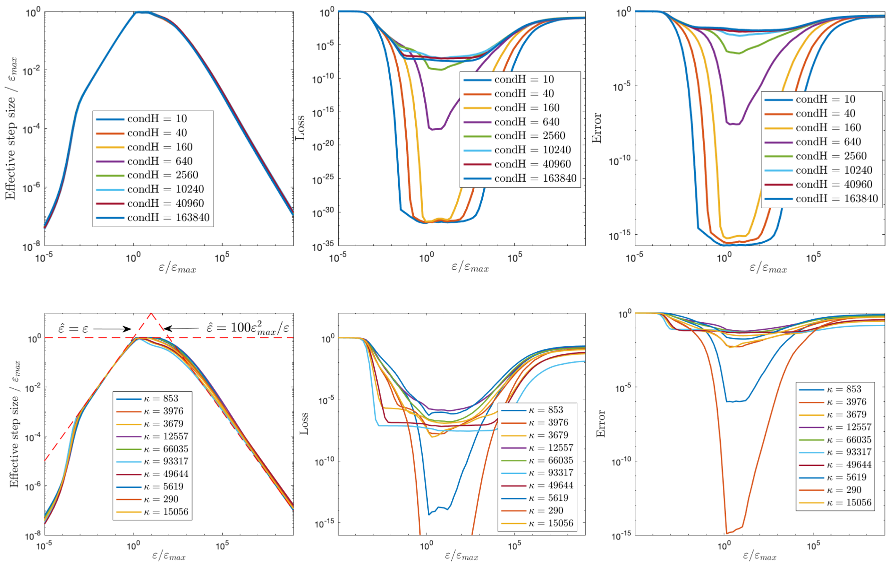

Figure 8. Tests of BNGD on OLS model with step size $\varepsilon_a = 1, a_0 = 0$. Parameters: (top) H =diag(linspace(1,condH,100)), (bottom) $H \in \mathbb{R}^{100\times 100}$ is a diagonal matrix with random positive entrances which has condition number κ. u and w_0 is randomly chosen uniformly from the unit sphere in $\mathbb{R}^{100}$. The BNGD iterations are executed for $k = 5000$ steps. The curves are averaged over the 500 independent runs.