Final presentation (EE5111 - Estimation Theory

Group-10

Gokul Mohanraj (EE17B106)
Shikhar Prakash (EE17B064)
Saurabh V (EP17B028)
Gayathri C (EP17B004)

Implicit Regularisation and Convergence in Weight Normalisation

- Explicit regularisation: Tikhonov, Dropout in Neural Networks.

$$J(\mathbf{x}) = \frac{1}{2}||\mathbf{y}\text{-}\mathbf{A}\mathbf{x}||^2 + \frac{\lambda}{2}||\mathbf{x}||^2$$

- Implicit Regularisation: Regularisation (implicitly) part of algorithm.
- Weight Normalisation.

$$J(\mathbf{x}) = H(g, \frac{\mathbf{w}}{||\mathbf{w}||}) \quad s.t.\, \mathbf{x} = g\frac{\mathbf{w}}{||\mathbf{w}||}$$

Weight Normalisation in Least Squares

$$L = \frac{1}{2}\|y - \frac{Agw}{\|w\|}\|^2$$

Algorithm 1 WN for $(\boxed{2})$

Input: Unit norm w_0 and scalar g_0,iterations T, step-sizes $\{\gamma_t\}_{t=0}^{T-1}$ and $\{\eta_t\}_{t=0}^{T-1}$

for $t = 0, 1, 2, \cdots, T-1$ **do**

$w_{t+1} = w_t - \eta_t \nabla_w h(w_t, g_t)$

$g_{t+1} = g_t - \gamma_t \nabla_g h(w_t, g_t)$

end for

Reparameterized Projected Gradient Descent(rPGD)

$$\min_{g\in\mathbb{R}, w\in\mathbb{R}^d} f(w,g) := \frac{1}{2}\|Agw - y\|^2, \text{ s.t. } \quad \|w\| = 1.$$

Algorithm 2 rPGD for (3)

Input: Unit norm w_0 and g_0, number of iterations T, step-sizes $\{\gamma_t\}_{t=0}^{T-1}$ and $\{\eta_t\}_{t=0}^{T-1}$

for $t = 0, 1, 2, \cdots, T-1$ **do**

$v_t = w_t - \eta_t \nabla_w f(w_t, g_t)$ (gradient step)

$w_{t+1} = \frac{v_t}{\|v_t\|}$ (projection)

$g_{t+1} = g_t - \gamma_t \nabla_g f(w_t, g_t)$ (gradient step)

end for

Lemma: Limiting Flow for WN and rPGD

Assumptions:

1. $\eta_t = \eta$ and $\gamma_t = c\eta$ with $c \geq 0$
2. $\|w_0\| = 1$

WN and rPGD have the same limiting dynamics or 'WN flow', that is:

$$\frac{dg_t}{dt} = -c\nabla_g f(w_t, g_t) \quad \frac{dw_t}{dt} = -g_t \mathcal{P}_t \left(\nabla_w f(w_t, g_t)\right)$$

where:

$\nabla_w f = A^T r$, $\nabla_g f = w^T A^T r$ and $r = y - Agw$ & $\mathcal{P}_t = I - w_t w_t^\top / \|w_t\|^2$

Proof (rPGD)

Let $a_t = \nabla_{w_t} f(w_t, g_t) = g\nabla L(gw)$

Expand the w_t update: $\|w_t - \eta\nabla_{w_t} f(w_t, g_t)\|_2^2 = \|w_t - \eta a_t\|_2^2 = 1 - 2\eta w_t^\top a_t + O(\eta^2)$

Now, $$\begin{aligned} w_{t+1} &= \frac{w_t - \eta\nabla_{w_t} f(w_t, g_t)}{\|w_t - \eta\nabla_{w_t} f(w_t, g_t)\|_2} \\ &= \frac{w_t - \eta a_t}{1 - \eta w_t^\top a_t + O(\eta^2)} \\ &= (w_t - \eta a_t) \cdot (1 + \eta w_t^\top a_t + O(\eta^2)) \\ &= w_t - \eta \mathcal{P}_t a_t + O(\eta^2). \end{aligned}$$

$$\boxed{\dot{w}_t = -g_t \mathcal{P}_t \nabla L(g_t w_t)}$$

Proof (WN)

Update steps:

$$v_t = w_t / \|w_t\|$$

$$g_{t+1} = g_t - c\eta \cdot \langle v_t, \nabla L\left(g_t \frac{w_t}{\|w_t\|}\right)\rangle$$

$$w_{t+1} = w_t - \eta \cdot g_t \cdot \mathcal{P}_t \frac{1}{\|w_t\|} \nabla L(g_t \frac{w_t}{\|w_t\|})$$

$$\dot{g}_t = -c \cdot \langle v_t, \nabla L(g_t v_t)\rangle$$

$$\dot{w}_t = -g_t \cdot \mathcal{P}_t \frac{1}{\|w_t\|} \nabla L(g_t v_t) = -\frac{g_t}{\|w_t\|} \cdot \mathcal{P}_t \nabla L(g_t v_t)$$

Now, $\frac{d\|w_t\|^2}{dt} = 2w_t^T \dot{w}_t = 0$ which gives $\|w_t\| = \|w_0\| = 1$

$$\boxed{\begin{aligned} \dot{g}_t &= -c \cdot w_t^T A^T r_t \\ \dot{w}_t &= -g_t \cdot \mathcal{P}_t A^T r_t \end{aligned}}$$

Lemma: Stationary Points

Suppose the smallest eigenvalue of AA^T , is positive, $\lambda_{min} := \lambda_{min}(AA^T) > 0$. The stationary points of the reparameterized loss either

(a) have loss equal to zero, or

(b) belong to the set $S := \{(g, w) : g = 0, y^T Aw = 0\}$.

Proof

- Similar to gradient descent proof. If loss = 0, we have reached the optimal w* and g*.

- Otherwise, we get w and g corresponding to the least squared solutions of the squared error loss.
  If g != 0, then from $\partial_g h(w,g) = w^T A^T r = 0$, $\partial_w h(w,g) = g \cdot P_{w^\perp} A^T r = 0.$ we get $P_{w^\perp} A^T r = 0.$ Using the first two equations and that ||w|| = 1, we can conclude $A^T r = 0$. Since $\lambda_{min} > 0$, r = 0 which means g != 0 is a zero-loss point with g = g* and w = w*.

  If g = 0, then solve y = (Aw)g algebraically for g. The least squares solution is $g = ((Aw)^T Aw)^{-1}(Aw)^T y = 0$ as g = 0. $((Aw)^T Aw)^{-1}$ is a non-zero scalar as $\lambda_{min} > 0$, hence $(Aw)^T y = 0$, or $y^T(Aw) = 0$. This is the set of points, S.

Rate of || r_t ||

Given the conditions of lemma 2.2, we have the following bound.

`

$$\begin{aligned} d[1/2\|r_t\|^2]/dt &= r_t^T \dot{r}_t = r_t^T Ad(g_t w_t)/dt \\ &= r_t^T A[\dot{g}_t w_t + g_t \dot{w}_t] \\ &= -r_t^T A[c \cdot w_t w_t^T A^T r_t + g_t^2 \mathcal{P}_t A^T r_t] \\ &= -r_t^T A[c \cdot w_t w_t^T + g_t^2 \mathcal{P}_t] A^T r_t \end{aligned}$$

$$d[1/2\|r_t\|^2]/dt \leq -\min\{g_t^2, c\}\|A^T r_t\|^2$$

$$\min(g_t^2, c) r_t^T A A^T r_t \geq \min(C^2, c)\lambda_{\min}(AA^T)\|r_t\|^2$$

and so with $k := \min(C^2, c)\lambda_{\min}(AA^T)$,

$$d[1/2 \cdot \|r_t\|^2]/dt \leq -k\|r_t\|^2 \quad \Rightarrow \quad \|r_t\|^2 \leq \exp(-kt)\|r_0\|^2$$

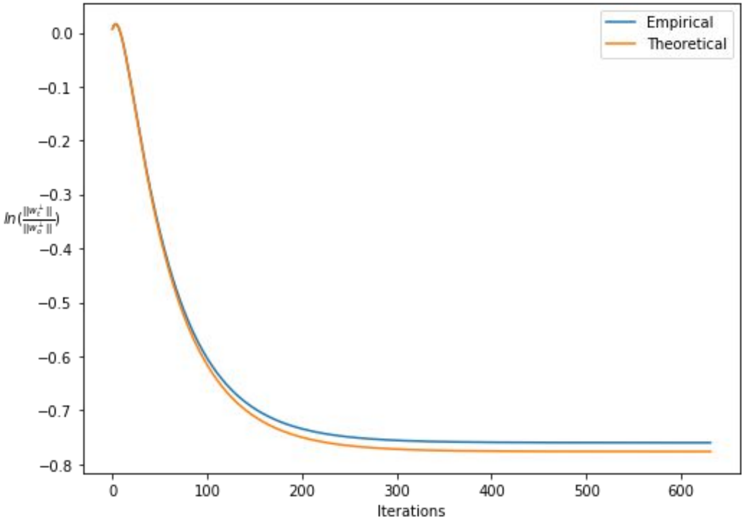
This shows that r_t is non-increasing and for some C > 0, g_t > C for all t, the loss decreases geometrically at the rate min(C^2, c).

Lemma: Orthogonal Component Variation

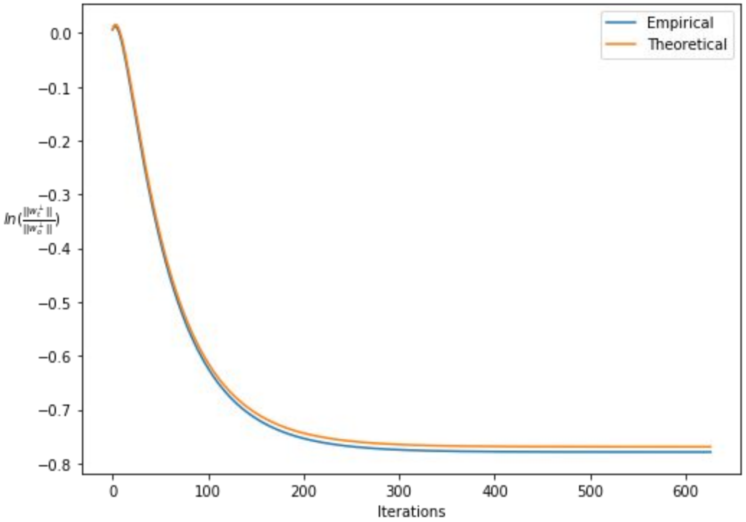
Assuming constant stepsizes for the updation of the weights as well as the scale parameters, with a constant ratio 'c' > 0, between the two step sizes, the following holds.

$$w_t^{\perp} = \exp\left(\frac{g_0^2 - g_t^2}{2c}\right) w_0^{\perp}$$

WN



rPGD



Proof and Observations

$$\frac{dw_t^{\perp}}{dt} = P^{\perp}\frac{dw_t}{dt} = -\frac{g_t}{\|w_t\|}P^{\perp}(I - \frac{w_t w_t^{\top}}{\|w_t\|^2})\nabla_w h_t$$

$$= \frac{g_t}{\|w_t\|}P^{\perp}\frac{w_t w_t^{\top}}{\|w_t\|^2}\nabla_w h_t = \frac{g_t}{\|w_t\|^2}P^{\perp}w_t\left(\nabla_g h_t\right) \quad = -\frac{g_t}{\|w_t\|^2}P^{\perp}w_t\left(\frac{1}{c}\frac{dg_t}{dt}\right)$$

$$\frac{dw_t^{\perp}}{dt} = -\frac{1}{2}\frac{w_t^{\perp}}{\|w_0\|^2 c}\frac{dg_t^2}{dt}$$

1. The orthogonal complement of w can change during the WN dynamics. This is the key property that yields the additional regularization.
2. It also suggests that $\|w_t^{\perp}\|^2 \cdot \exp(g_t^2/2c)$ remains invariant along the WN path.

Weight Normalization Flow Solution

Assuming constant stepsizes for the updation of the weights as well as the scale parameters, with constants c, η, g and w defined as before, suppose the flow is initialized at g_o and w_o where $||w_o|| = 1$, either one of the following holds:

(a) The loss converges to zero

(b) iterates (g_t, w_t) converge to a stationary point in S as defined in Lemma 2.3.

If (a) holds, then we characterize the solutions based on g_t as follows:

Part I: If c>0, $\lim_{t\to\infty} g_t w_t = x^* + g^* w_0^\perp \exp\left(\frac{g_0^2 - g^{*2}}{2c}\right)$ *and sufficient condition for convergence is* $\|y\|^2 > \|Ag_0w_0 - y\|^2$

Part II: If c = 0, and A is orthogonal, then $w_t \to w^*$. If A is not orthogonal, then the flow still converges to a point $\tilde{w}_o$ in the row space of A. When restarting the WN flow with c > 0 from g_o, $\tilde{w}_o$, then $(g_o, \tilde{w}_o) \to (g^*, w^*)$.

Proof

(a) From lemma 2.4: $d[1/2\|r_t\|^2]/dt = -r_t^T A[c \cdot w_t w_t^T + g_t^2 \mathcal{P}_t] A^T r_t$. If g_t^2 -> C^2 and C > 0, then the loss converges to zero according to lemma 2.4.

Hence, we focus on the case when C = 0. Now, if g_t -> 0, we have x_t -> 0. Then,

$$\|r_t\|^2 = \|y - Ag_t w_t\|^2 \to \|y\|^2.$$

(b) Suppose on the contrary that the loss does not converge to 0, then $||r_t||$ -> c" for some constant c" > 0. Given g_t -> 0, it can be shown from lemma 2.4 that: $(r_t^T A w_t)^2 \to 0.$ Otherwise, $d[1/2\|r_t\|^2]/dt < -c'$ for c' > 0 and we'll have unbounded decrease of $||r_t||$, a contradiction. Thus, if $||r_t||$ does not converge to 0, then x_t -> 0 since g_t -> 0. Also, y^TAw_t -> 0 from lemma 2.3. Hence, (g_t, w_t) converge to a set S defined as:

$$S := \{(g, w) : g = 0, y^T Aw = 0\}.$$

Explanations for part (a) and (b)

Part I: From lemma 2.5, the orthogonal component converges to the invariant form.
Row space component converges to $w_{||}^*$ and g converges to g*. Let $x^* = g^* w_{||}^*$. This gives the convergence of $x_t = g_t^* w_t$ as required:

$$\lim_{t\to\infty} g_t w_t = x^* + g^* w_0^{\perp} \exp\left(\frac{g_0^2 - g^{*2}}{2c}\right)$$

Part II:

For orthogonal A, even fixing the scale g_o we can converge to the direction of the minimum norm solution. Hence, once we have decided the correct direction of w, g* can be recovered as $|g^*| = ||y||$.

For general A with fixed g, we do not necessarily converge to the right direction w*, only to the row span of A. Hence, run the flow with c = 0 until convergence, and then turn on the flow for g (i.e. set c > 0), to get the minimum norm solution.

Refer to the slide in appendix for more details

Observations from the Theorem

- This theorem gives us a proof that either loss converges to zero or it converges to a stationary point as defined in lemma 2.3, with $g^* = 0$.
- The case when the loss converges to 0, the flow dynamics and the WN solution are determined by update rule of g_t the orthogonality of A.
- If g_t is updated, then the solution has a definite form.
- If g_t is not updated, then if A is orthogonal, then w -> w^* and g^* can be determined numerically ($||g^*|| = ||y||$).
- If g_t is not updated and if A is not orthogonal, then w -> $\tilde{w}$. Now, restart WN Flow dynamics from this w in order to reach the optimal direction, w^*.
- Does not give rate of convergence.

Theorem 2.7

Assumptions:

1. $\eta_t = \eta$ and $\gamma_t = c\eta$ with $c \geq 0$
2. $\|w_0\| = 1$
3. Smallest eigenvalue $\lambda_{\min}$ of AA^T is strictly positive

The loss decreases and $f(w_T, g_T) \leq \varepsilon$ after time T

If $g_0^2 > 2c\log(1/\|w_0^\perp\|)$: $$T = \frac{\log(f(w_0, g_0)/\varepsilon)}{\lambda_{\min}\min\left\{2c\log\|w_0^\perp\| + g_0^2, c\right\}}$$

$\delta = (\|y\|^2 - \|Ag_0w_0 - y\|^2)/\lambda_{\max} > 0$; $$T = \frac{\log(f(w_0, g_0)/\varepsilon)}{\lambda_{\min}\min\{\delta, c\}} + \frac{1}{\lambda_{\max}}\log\left(2 - \frac{g_0}{\delta}\right)\mathbb{1}_{\{g_0 < \delta\}}$$

Proof (Case 1)

We have, $g_t^2 = 2\log\|w_0^\perp\| + g_0^2 - 2\log\|w_t^\perp\| \geq 2\log\|w_0^\perp\| + g_0^2$

Because, $\|w_t^\perp\| \leq \|w_t\| = 1$

Therefore, loss decreases geometrically at a rate given by the minimum of the lower bound obtained above and c. Hence,

$$T = \frac{\log(f(w_0, g_0)/\varepsilon)}{\lambda_{\min} \min\left\{2c\log\|w_0^\perp\| + g_0^2, c\right\}}$$

Proof (Case 2)

Lemma C.1:

Assumptions: $\frac{\|Ag^*w^*\|^2 - \|A(g_0w_0 - g^*w^*)\|^2}{\lambda_{\max}} > \delta$ for some small δ

Lower bound: $g_t \geq \frac{\left(\|Ag^*w^*\|^2 - \|A(g_0w_0 - g^*w^*)\|^2\right)}{\lambda_{\max}} - \frac{\delta}{2}$ *for* $t \geq s$

where,

$$s = \begin{cases} 0 & \textit{if } g_0 \geq \min\left\{\frac{2g^*\langle Aw_0, Aw^*\rangle}{\|Aw_0\|^2}, \frac{\left(\|Ag^*w^*\|^2 - \|A(g_0w_0 - g^*w^*)\|^2\right)}{\lambda_{\max}} - \delta\right\} \\ \frac{1}{\lambda_{\max}} \log\left(\frac{2}{\delta}\left(\frac{\|Ag^*w^*\|^2 - \|A(g_0w_0 - g^*w^*)\|^2}{\lambda_{\max}} - g_0\right)\right) & \textit{otherwise.} \end{cases}$$

$$\boxed{T = \frac{\log(f(w_0, g_0)/\varepsilon)}{\lambda_{\min}\min\{\delta, c\}} + \frac{1}{\lambda_{\max}} \log\left(2 - \frac{g_0}{\delta}\right) \mathbb{1}_{\{g_0 < \delta\}}}$$

Orthogonal Matrix Convergence

Theorem 3.2 (Convergence for Orthogonal Matrix A). *Suppose the initialization satisfies $0 < g_0 < g^*$, and that w_0 is a vector with $\|w_0\| = 1$. Let $\delta_0 = (g^*)^2 - (g_0)^2$. Set an error parameter $\varepsilon > 0$ and the stepsize given in Condition 3.1 with a hyper-parameter $\rho \in (0, 1]$ for $\gamma^{(1)}$. Running the rPGD algorithm, we can reach $\|w_{T_1}^{\perp}\| \le \varepsilon$ and $g_{T_1}^2 \le g^{*2} - \rho\delta_0$ after T_1 iterations, and $\|w_T^{\perp}\| \le \varepsilon$ and $\|A g_T w_T - b\|^2 \le 3\varepsilon g^{*2}$ after $T = T_1 + T_2$ iterations, if we set stepsizes as follows.*

(a) Set $\gamma^{(1)} = \mathcal{O}\left(\frac{\rho}{\log(1/\varepsilon)}\left(\frac{g_0}{g^*}\right)^2 \log\left((1-\rho)\frac{g^*}{g_0} + \rho\right)\right)$ and $\gamma^{(2)} \le \frac{1}{4}$. Then we have

$$T_1 = \mathcal{O}\left(\frac{(g^*)^2}{\rho\delta_0}\log\left(\frac{1}{\varepsilon}\right)\right); \quad T_2 = \mathcal{O}\left(\frac{1}{\gamma^{(2)}}\log\left(\frac{(\rho\delta_0/g^{*2})^2}{\varepsilon}\right)\right).$$

(b) Set $\gamma^{(1)} = 0$ and $\gamma^{(2)} < \frac{1}{4}$. Then we have

$$T_1 = \mathcal{O}\left(\frac{g_0^2}{\delta_0}\log\left(\frac{1}{\varepsilon}\right)\right); \quad T_2 = \mathcal{O}\left(\frac{1}{\gamma^{(2)}}\log\left(\frac{\sqrt{\delta_0/g^{*2}}}{\varepsilon}\right)\right).$$

Gist of the proof (conditions)

$$
\begin{aligned}
v_t &\overset{(a)}{=} w_t - \eta_t g_t^2 A^\top A w_t + \eta_t g_t g^* A^\top A w^* \\
&\overset{(b)}{=} (I - A^\top A) w_t + \frac{g^*}{g_0} A^\top A w^* \\
&\overset{(c)}{=} w_t^\perp + \frac{g^*}{g_0} w^*,
\end{aligned}
$$

$$
\|w_{t+1}^\perp\|^2 = \frac{\|v_t^\perp\|^2}{\|v_t\|^2} = \frac{\|w_t^\perp\|^2}{\|w_t^\perp\|^2 + g^{*2}/g_0^2} \le \frac{g_0^2}{g^{*2}} \|w_t^\perp\|^2.
$$

Since $g_0 < g^*$, after $T_1 = \frac{\log(1/\varepsilon^2)}{\log(g^{*2}/g_0^2)}$ iterations, we have

$$
\|w_{T_1}^\perp\|^2 \le (g_0^2/g^{*2})^{T_1} \le \varepsilon^2.
$$

$$
\begin{aligned}
g_{t+1} &= g_t - \gamma g_t w_t^T A^\top A w_t + \gamma g^* w_t^T A^\top A w^* \\
&\overset{(a)}{=} g_t - \gamma g_t \|w_t^\parallel\|^2 + \gamma g^* \left\langle w_t^\parallel, w^* \right\rangle, \\
&\overset{(b)}{=} g_t - \gamma g_t \|w_t^\parallel\|^2 + \gamma g^* \|w_t^\parallel\|,
\end{aligned}
$$

- Property (i): $\|w_{t+1}^\perp\| \leq \|w_t^\perp\| \leq \varepsilon$.
- Property (ii): letting $\gamma' = \gamma(1-\varepsilon^2)$, we have

$$
(1-\gamma')g_t + \gamma' g^* \leq g_{t+1} \leq g
$$

$$
\begin{aligned}
g^* - g_T &\leq (1-\gamma')(g^* - g_{T-1}) \\
&\leq (1-\gamma')^{T_2}(g^* - g_{T_1}) \\
&\overset{(a)}{=} (1-\gamma')^{T_2}(g^* - g_0) \\
&\overset{(b)}{\leq} 2\varepsilon^2 g^*,
\end{aligned}
$$

$$
\begin{aligned}
f(w_T, g_T) &= g_T^2 \|Aw_T\|^2/2 - g_T g^* \langle Aw_T, Aw^* \rangle + g^{*2}/2 \\
&\leq g^{*2}/2 - (1-2\varepsilon^2) g^{*2} (1-\varepsilon) + g^{*2}/2 \\
&\leq 3\varepsilon g^{*2}.
\end{aligned}
$$

Lemma E.6. *We have the following bound on the closeness of Aw_t to unit norm:*

$$\|w_t^\perp\| \leq (1 - \|Aw_t\|^2) \leq \exp(-\sum_{i=1}^{t} \frac{(g^*)^2 - \|Ag_i w_i\|^2}{(g^*)^2 + (g^*)^2 - \|Ag_i w_i\|^2})(1 - \|Aw_0\|^2$$

$$\begin{aligned}
1 - \|Aw_{t+1}\|^2 =& \frac{g_t^2(1 - \|Aw_t\|^2)}{(g^*)^2 + g_t^2(1 - \|Aw_t\|^2)} \\
\leq& \frac{(g^*)^2}{(g^*)^2 + (g^*)^2 - \|Ag_t w_t\|^2}(1 - \|Aw_t\|^2) \\
\leq& \exp(-\frac{(g^*)^2 - \|Ag_t w_t\|^2}{(g^*)^2 + (g^*)^2 - \|Ag_t w_t\|^2})(1 - \|Aw_0\|^2).
\end{aligned}$$

Thus,

$$(1 - \|Aw_t\|^2) \leq \exp(-\sum_{i=1}^{t} \frac{(g^*)^2 - \|Ag_i w_i\|^2}{(g^*)^2 + (g^*)^2 - \|Ag_i w_i\|^2})(1 - \|Aw_0\|^2).$$

$$(g^*)^2 - \|Ag_{T_1}w_{T_1}\|^2 \geq (g^*)^2 - g_{T_1}^2 = \rho\delta_0$$

By Lemma E.6, we have

$$\|w_{T_1}^{\perp}\|^2 = (1 - \|Aw_{T_1}\|^2) \leq \exp(-\sum_{i=1}^{T_1} \frac{(g^*)^2 - \|Ag_iw_i\|^2}{(g^*)^2 + (g^*)^2 - \|Ag_iw_i\|^2})(1 - \|Aw_0\|^2)$$

$$\leq \exp(-\sum_{i=1}^{T_1} \frac{(g^*)^2 - \|Ag_{T_1}w_{T_1}\|^2}{(g^*)^2 + (g^*)^2 - \|Ag_{T_1}w_{T_1}\|^2})(1 - \|Aw_0\|^2)$$
$$\leq \exp(-\frac{\rho\delta_0 T_1}{(g^*)^2 + \rho\delta_0})(1 - \|Aw_0\|^2)$$

we have $\|w_{T_1}^{\perp}\|^2 = 1 - \|Aw_{T_1}\|^2 \leq \delta^2$ when

$$T_1 = \left(1 + \frac{(g^*)^2}{\rho\delta_0}\right) \log\left(\frac{1 - \|Aw_0\|^2}{\delta^2}\right)$$

Observations from the Theorem

1. This shows that the rPGD converges to the minimum norm solution at the rate $\log(1/\varepsilon)$ when all others are constants. The first T1 iterations allow the algorithm to find w^* and the remaining T2 to find g^*.
2. There is an intrinsic tradeoff. Larger δ_o allows results in smaller T1 but a larger time for convergence of g_t.
3. Also we find that $w^{\perp}_t$ decreases at a geometric rate.
4. Lastly, when A is orthogonal, for the optimal stepsize, we can escape the saddle points and reach the global minimum.

Number of Iterations Needed for Convergence

Fix δ > 0, and fix a full rank matrix A with $\lambda_{max}(AA^T) = 1$. With a fixed $g = g_o$ satisfying $g_o \leq [g^*\lambda_{min}(AA^T)]/(2+\delta)$, we can reach a solution with $// w^{\perp} //\ \leq \varepsilon$ in a number of iterations given as:

$$T_1 = \log\left(\frac{\|w_0^{\perp}\|}{\varepsilon}\right) / \log(1+\delta).$$

Proof

A proof of the weaker version of conditions of this theorem for rPGD is easier to prove. Consider g_o satisfies $g_0 \leq \frac{g^* \sigma_r}{2 + \delta - \sigma_r}$ where r is the rank of A, $\sigma_m = \lambda_{min}(AA^T)$ with m <= r and σ_i are the singular values of A in decreasing order.

Consider singular value decomposition of $A^TA = U\Sigma U^T$. Here, U is a dxd orthogonal matrix and Σ is given by:

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_m & \\ & & & \mathbf{0}_{d-m} \end{bmatrix} \quad \text{with } 1 = \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m > 0.$$

Now with $\lambda_{max}(AA^T) = \sigma_1 = 1$, let $\eta = 1/(g_t^2 \sigma_1) = 1/(g_t^2)$. The update for v_t is given by:

Recall for rPGD $v_t = w_t - \eta_t \nabla_w f(w_t, g_t)$ (gradient step)

Proof

Update for v_t is: $v_t = w_t - \eta g_0 A^\top A(g_0 w_t - g^* w^*) = (I - A^\top A)w_t + \frac{g^*}{g_0} A^\top A w^* = U\left(I - \Sigma\right)U^\top w_t + \frac{g^*}{g_0} U \Sigma U^\top w^*$

$$
\begin{aligned}
\|v_t\| &= \|\frac{g^*}{g_0}\Sigma U^\top w^* + \left(I - \Sigma\right)U^\top w_t\| \\
&= \left\| \frac{g^*}{g_0} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_m & \\ & & & \mathbf{0}_{d-m} \end{bmatrix} U^\top w^* + \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \mathbf{1}_{d-m} \end{bmatrix} U^\top w_t + \begin{bmatrix} 1-\sigma_1 & & & \\ & \ddots & & \\ & & 1-\sigma_m & \\ & & & \mathbf{0}_{d-m} \end{bmatrix} U^\top w_t \right\| \\
&\geq \sqrt{\left(\frac{g^*}{g_0}\right)^2 \sum_{i=1}^{m} \sigma_i^2 [U^\top w^*]_i^2 + \sum_{i=m+1}^{d} [U^\top w_t]_i^2} - \sqrt{\sum_{i=1}^{m} (1-\sigma_i)^2 [U^\top w_t]_i^2} \\
&\geq \frac{g^*}{g_0}\sigma_m - (1 - \sigma_m) \\
&\geq \left(\frac{g^*}{g_0} + 1\right)\sigma_m - 1 \\
&\geq 1 + \delta
\end{aligned}
$$

Now, $\sigma_m \leq \sigma_1 = 1$, the following holds: $\sigma_m \geq \dfrac{2+\delta}{\left(\frac{g^*}{g_0} + 1\right)} \quad \Leftrightarrow \quad g_0 \leq \dfrac{g^* \sigma_m}{2 + \delta - \sigma_m}$ and $\sigma_m \leq 2$

Proof

The inequalities give that as long as g_o is small, $||v|| \geq 1 + \delta$.

Hence, by definition, we get

$$\|w_{t+1}^{\perp}\| = \frac{\|w_t^{\perp}\|}{\|v_{t+1}\|} \leq \frac{1}{1+\delta}\|w_t^{\perp}\|$$

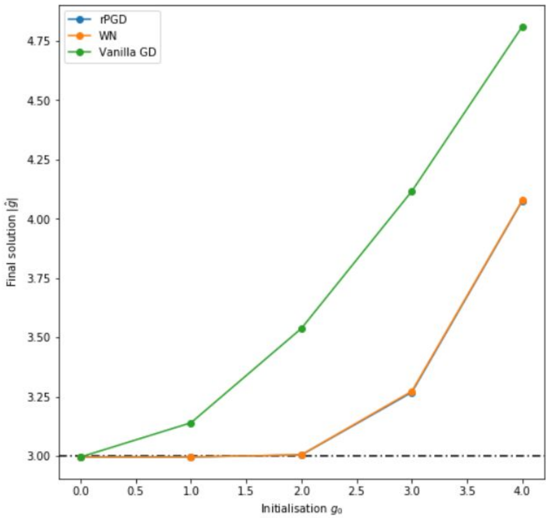
Which iteratively gives: $$||w_{T_1}^{\perp}|| \leq \frac{1}{(1+\delta)^{T_1}}||w_0^{\perp}||$$

Now, since $||w_{T_1}^{\perp}|| \leq \epsilon$

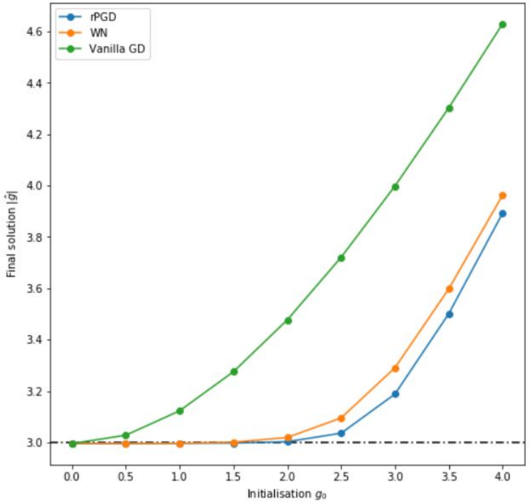
We solve for T_1 to get the desired result:

$$T_1 = \frac{1}{\log(1+\delta)} \log\left(\frac{\|w_0^{\perp}\|}{\varepsilon}\right).$$

Final solution g^* vs initialisation g_0



GD, WN and rPGD in continuous time (0.005 step size)



GD, WN and rPGD for discrete time (γ = η = 0.1)

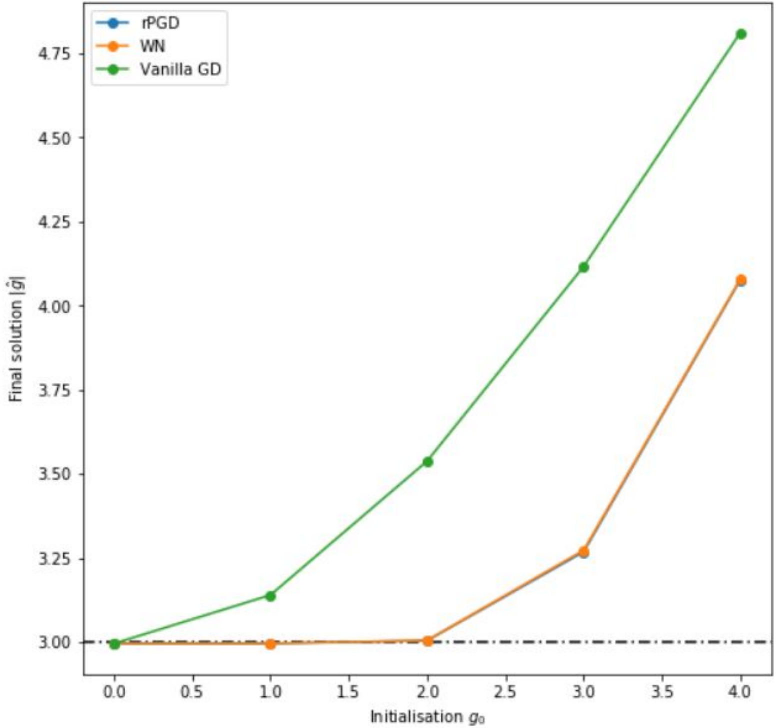
Continuous case:

- Both WN and rPGD perform better than GD
- WN and rPGD have close flows

Discrete Case:

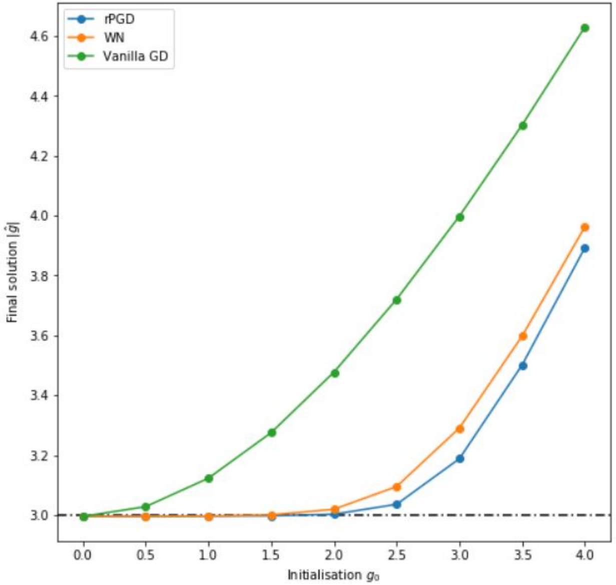
- rPGD and WN outperform vanilla GD for larger values of initialization
- rPGD slightly better than WN

GD, WN and rPGD in continuous time (0.005 step size)

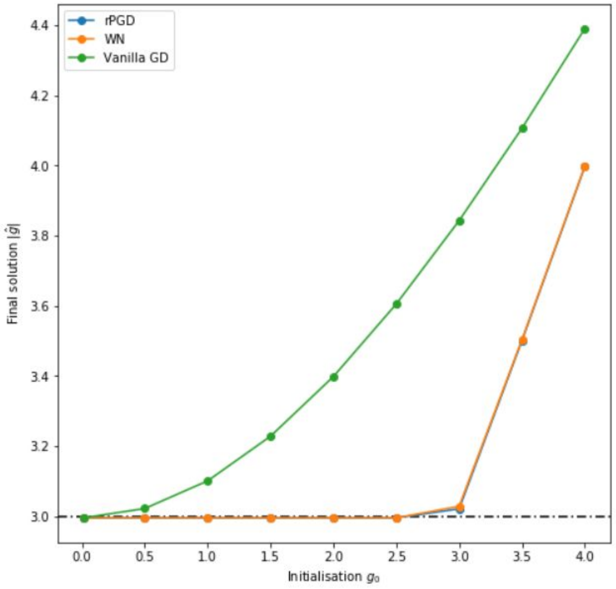


1. Vanilla GD does not converge to true norm when initialized far from origin.
2. WN and rPGD converge to true norm with bigger range of initializations

GD, WN, rPGD for Discrete Time and 2 Phase



$\gamma = \eta = 0.1$ (Discrete case)

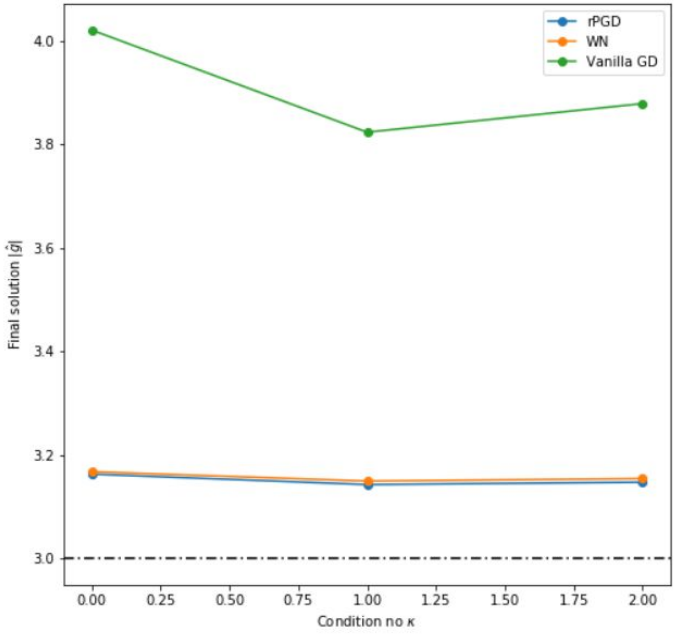


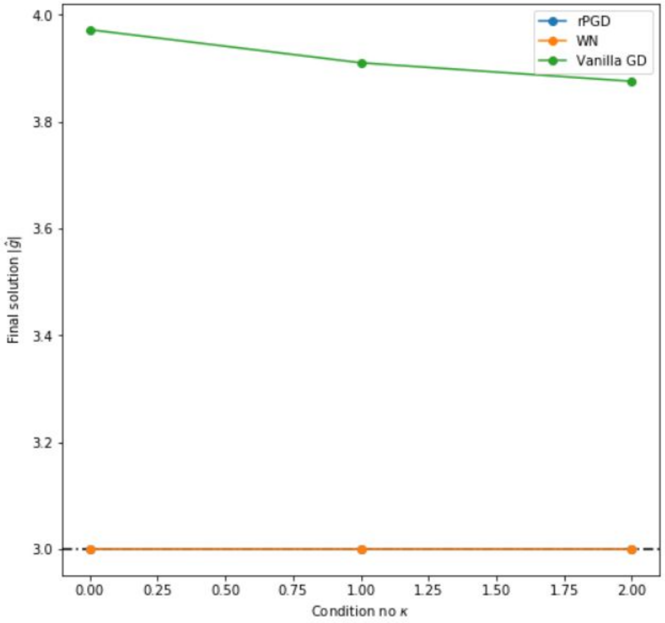
Two phase implementation

1. Both rPGD and WN outperform vanilla GD for larger values of initialization

2. rPGD slightly better than WN

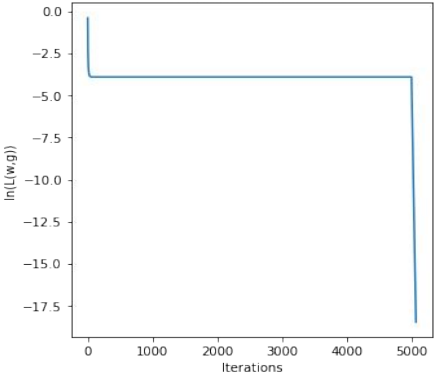
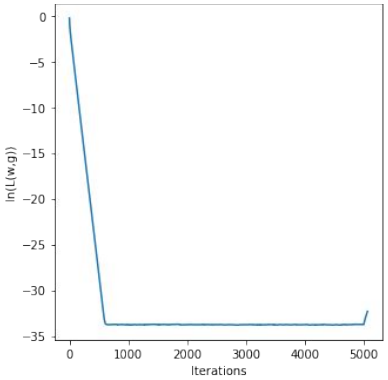
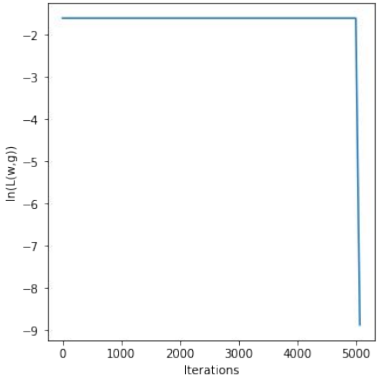
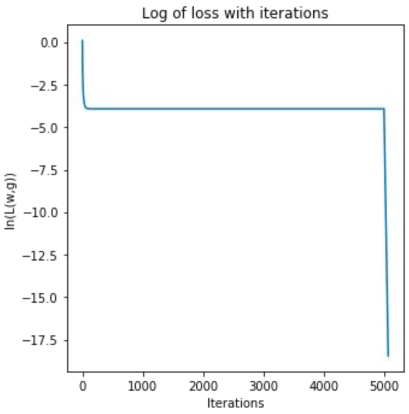
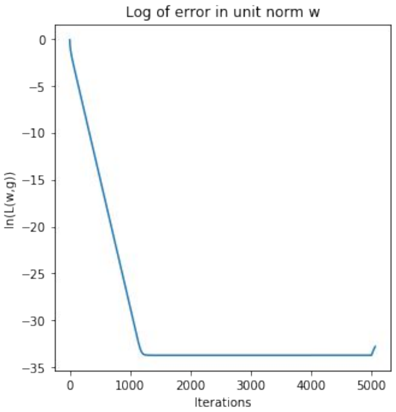
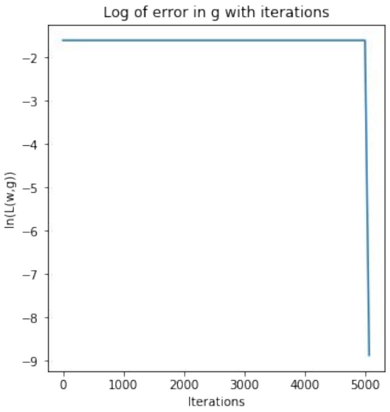
GD, WN, rPGD under Varying Condition Number





1. Both rPGD and WN outperform vanilla GD.
2. In the two phase case, both rPGD and WN converge exactly to the minimum norm solution.

Dynamics of Two Phase Algorithm

	Log of loss with iterations	Log of error in unit norm w	Log of error in g with iterations
rPGD			
WN			

THANK YOU

Appendix: Explanation for part (b) of WN Flow Solution

We have:

$$
\begin{aligned}
d[1/2 \cdot \|r_t\|^2]/dt &= r_t^T \dot{r}_t = r_t^T A d(g_t w_t)/dt \\
&= r_t^T A[\dot{g}_t w_t + g_t \dot{w}_t] \\
&= r_t^T A g_0 \dot{w}_t \\
&= -r_t^T A g_0 \mathcal{P}_t g_0 A^T r_t \\
&= -g_0^2 \|\mathcal{P}_t A^T r_t\|^2.
\end{aligned}
$$

Solving for r_t and then for w_t since $r_t = y - Ag_t w_t$, we get $w_t = \pm A^T r/||A^T r||$. With $||w|| = 1$, only the $w=-A^T r/||A^T r||$ minimizes the loss. The optimal solution is $w=-A^{+-} r/||A^{+-} r||$. Here, A^{+-} is the pseudoinverse of A. If A is orthogonal, $A^{+-} = A^T$ and this $w = w^*$.

Otherwise, consider:

$$
\begin{aligned}
d[1/2 \cdot \|r_t\|^2]/dt &= r_t^T \dot{r}_t = r_t^T A d(g_t w_t)/dt \\
&= r_t^T A[\dot{g}_t w_t + g_t \dot{w}_t] \\
&= -r_t^T A[w_t w_t^T A^T r_t + g_t^2 \mathcal{P}_t A^T r_t] \\
&= -(r_t^T A w_t)^2.
\end{aligned}
$$

Again, $r_t^T A w_t \rightarrow 0$ under limit $t \rightarrow \infty$. Hence, solve for r_t and then for w_t since $r_t = y - Ag_t w_t$ with w at t = 0 to be $w=-A^T r/||A^T r||$. Keeping $||w|| = 1$, we achieve $w = w^*$.