We thanks the reviewers for their feedback. We now proceed to reply to their comments.

Number of Samples for Complete Graph? (R1): In this case there is no network error and all of the agents iterates are identical to centralised single-machine Gradient Descent (GD) applied to all of the samples within the network, as we point out in Appendix B.1. Error Decomposition. There is no condition on the number of samples to achieve optimal statistical rates in this case [25].

Extending Current Limitations (R2, R3):
In light of the comments we have pursed extending the theory on the following fronts:

- **Non-attainable case**: Consider the capacity independent ($\gamma = 1$) non-attainable case ($r < 1/2$), when agents know the population covarince $\mathcal{T}_\rho$ i.e. additive noise oracle or fixed-design regression [15,20]. The network error in this case only consists of the population covariance error and can be bounded as follows. Due to being in the non-attainable case, the bound on the terms $\{N_{k,w_k}\}_{k,w_k}$ now depend on the time step k, in particular with regularisation becoming $\widetilde{O}\big(\frac{1}{\sqrt{m\lambda}} + \frac{(\eta k)^{1/2-r}}{m\sqrt{\lambda}}\big)$, see Lemma 18 in [26]. The first term matches the attainable case while the second is new and now must be controlled. Following the analysis for the attainable case, this new term squared can be shown to be $\widetilde{O}\big((\eta t)^{1-2r}(\eta t^\star)/m^2\big)$ in high probability. With $\eta = O(1)$ and $t = (nm)^{1/(2r+\gamma)}$ this quantity is smaller than the optimal statistical rate once $m \geq n^{1/(4r+1)}(t^\star)^{(2r+1)/(4r+1)}$. This condition is strictly weaker than $m \geq n^{2r}(t^\star)^{2r+1}$, the condition that arises in the attainable case to achieve a linear speed up by ensuring the first term is sufficiently small, if $r \geq 1/4$. We leave the capacity dependent non-attainable case to future work.
- **Residual Covariance Error**: When agents do not know the population covarince, we require $2r + \gamma \geq 2$ to control the residual covariance error. A consequence in the finite low-dimensional setting ($d < nm$,$\gamma = 0$) is that bounds only hold for "easy" problems $r > 1$. We note a similar consequence occurs in the the high-dimensional setting as source and capacity assumptions analogous to our setting can be utilised to achieve dimension free bounds (when $d > nm$ the classical $O(d/nm)$ rate is vacuous), see [34]. The limitation $2r + \gamma \geq 2$ can be improved though sharper control of the residual network error terms, in particular by repeatedly applying the recursion (Proposition 5) and centering the difference $\Pi_{t:k+1}(\mathcal{T}_{\mathbf{x}_{w_{t:k+1}}}) - \Pi_{t:k}(\mathcal{T}_\rho)$ around the expectation $\mathbf{E}[\Pi_{t:k+1}(\mathcal{T}_{\mathbf{x}_{w_{t:k+1}}})]$.

Communication Model (R3):

- **Description**: We consider a lockstep communication model where each round lasts for τ units of time. Within each round, agents send/receive the messages to/from their neighbours in order to implement a single update of algorithm (3). With a gradient evaluation costing 1 unit of time, each iteration of Distributed Gradient Descent takes the following amount of time $m + \tau + \mathrm{Deg}(P)$: m gradient evaluations; τ in communication delay ; $\mathrm{Deg}(P)$ for each agent to aggregating their neighbours and own gradients, as the sum in algorithm (3) $\sum_{w \in V} P_{vw}$ has computational cost $O(\mathrm{Deg}(P))$.
- **Communication Delay τ Increasing with $\mathrm{Deg}(P)$?**: Indeed τ maybe increasing with the network degree, although it could depend on other factors arising from: noisy transmission, compressing or decompressing messages and synchronizing with neighbours. The work Tsianos and Rabbat, NIPS 2012 makes the assumption that the delay τ is a linear function of the network degree and transmit time $r \geq 0$ so $\tau = r\mathrm{Deg}(P)$. The conclusion of Section 3.1 would be unaffected provided $\tau + \mathrm{Deg}(P) = O(m)$.
- **Tradeoff In Terms of the Parameter τ**: For sufficiently large m the speed up is $nm/(m + \tau + \mathrm{Deg}(P))$. No speed up is achieved if the communication delay is lower bounded $\tau = \Omega(nm)$, that is if it is longer than the time it takes for centralised single-machine GD to compute a gradient with all of the samples within the network. More precisely, suppose a network topology with $\mathrm{Deg}(P) = \Theta(n^\beta)$ for $0 \leq \beta \leq 1$, delay $\tau = r\mathrm{Deg}(P)$ as described previously and $m > n$. Then a linear speed up is achieved if $r \leq m/n^\beta$, no speed up if $r \geq mn^{1-\beta}$, and a non-linear speed up if $m/n^\beta < r < mn^{1-\beta}$. In contrast, for high degree graphs with $\beta = 1$ and $(1 - \sigma_2)^{-1} = O(1)$, Sianos and Rabbat, 2012 saw a linear speed up when $r < 1$ whilst we require $r \leq m/n$. We thank the reviewer for this question as the above is now included within the manuscript.

Communicating Infinite-Dimensional Quantities (R3): A theoretical setting allows precise understanding of the generalisation capabilties through optimal statistical rates [12]. While particular implementations are outside the scope of this work, note it is common to use finite approximations of infinite dimensional quantities whilst accounting for the statistical precision (Learning with sgd and random features Luigi, Rudi & Roasasco 2018).

Regime Where Nodes Work Independently (R3) : The objective is for each agent to achieve the optimal statistical rate with respect to *all* nm data points within the network, and as such, it is a requirement for agents to communicate one another.