GRAPH-HOMOMORPHIC PERTURBATIONS FOR PRIVATE DECENTRALIZED LEARNING

Stefan Vlaski and Ali H. Sayed

School of Engineering, École Polytechnique Fédérale de Lausanne

ABSTRACT

Decentralized algorithms for stochastic optimization and learning rely on the diffusion of information through repeated local exchanges of intermediate estimates. Such structures are particularly appealing in situations where agents may be hesitant to share raw data due to privacy concerns. Nevertheless, in the absence of additional privacy-preserving mechanisms, the exchange of local estimates, which are generated based on private data can allow for the inference of the data itself. The most common mechanism for guaranteeing privacy is the addition of perturbations to local estimates before broadcasting. These perturbations are generally chosen independently at every agent, resulting in a significant performance loss. We propose an alternative scheme, which constructs perturbations according to a particular nullspace condition, allowing them to be invisible (to first order in the step-size) to the network centroid, while preserving privacy guarantees. The analysis allows for general nonconvex loss functions, and is hence applicable to a large number of machine learning and signal processing problems, including deep learning.

Index Terms— Decentralized optimization, learning, differential privacy, encryption, diffusion strategy.

1. INTRODUCTION AND RELATED WORKS

We consider a collection of K agents, where each agent k is equipped with a local loss function:

$$J_k(w) \triangleq \mathbb{E}\, Q(w; \boldsymbol{x}_k) \tag{1}$$

The agents are interested in pursuing a minimizer to the aggregate optimization problem:

$$\min_w J(w) \triangleq \min_w \frac{1}{K}\sum_{k=1}^{K} J_k(w) \tag{2}$$

While the minimizer of (2) can be pursued by a variety of decentralized strategies, we focus here on the Adapt-Then-Combine (ATC) diffusion strategy due to its enhanced performance in adaptive scenarios [1]:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \nabla Q_k(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i}) \tag{3}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\phi}_{l,i} \tag{4}$$

The diffusion strategy (3)–(4) has strong performance guarantees in both the (strongly) convex [1–3] and non-convex [4,5] settings. For (3)–(4) to minimize (2), we will assume the combination weights to be symmetric and stochastic, i.e.:

$$a_{\ell k} = a_{k\ell}, \quad \sum_{\ell=1}^{K} a_{\ell k} = 1, \quad a_{\ell k} \geq 0 \tag{5}$$

When agents are concerned about privacy, they may be hesitant to share their intermediate estimates $\boldsymbol{\phi}_{k,i}$, since they can contain significant information about their locally observed data $\boldsymbol{x}_{k,i}$ through its evolution via the gradient. To see that this is the case, consider the least-squares loss $Q_k(w; \boldsymbol{h}, \boldsymbol{\gamma}) \triangleq \|\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}} w^o\|^2$, with stochastic gradient:

$$\nabla Q(w; \boldsymbol{h}, \boldsymbol{\gamma}) = \boldsymbol{h}\left(\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}} w\right) \propto \boldsymbol{h} \tag{6}$$

In other words, the stochastic gradient $\nabla Q(w; \boldsymbol{h}, \boldsymbol{\gamma})$ reveals the raw feature $\boldsymbol{h}$ up to a normalizing factor, and hence observation of iterates, which evolve according to $\nabla Q(w; \boldsymbol{h}, \boldsymbol{\gamma})$, allows for the inference of $\boldsymbol{h}$. A common strategy to ensure privacy in recursive algorithms is to perturb intermediate estimates before sharing them, resulting in [6]:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \nabla Q_k(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i}) \tag{7}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \boldsymbol{q}_{k,i} \tag{8}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\psi}_{l,i} \tag{9}$$

The added perturbation $\boldsymbol{q}_{k,i}$ is typically chosen to follow some zero-mean Gaussian or Laplacian distribution and essentially masks the gradient $\nabla Q_k(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i})$, which contains information about the data $\boldsymbol{x}_{k,i}$. This results in rigorous privacy guarantees (quantified by differential privacy [7]), but comes at a cost, namely non-negligible degradation in performance. To see why this is the case, we introduce the gradient noise:

$$\boldsymbol{s}_{k,i} \triangleq \nabla Q_k(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \tag{10}$$

Emails:{stefan.vlaski, ali.sayed}@epfl.ch.

Under this definition, recursions (7)–(9) can be written equivalently as:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \nabla J_k(\boldsymbol{w}_{k,i-1}) - \mu\, \boldsymbol{s}_{k,i} \tag{11}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i-1} + \boldsymbol{q}_{k,i} \tag{12}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\psi}_{l,i} \tag{13}$$

Inspection of (11)–(13) shows that the effect of privatizing the gradient $\nabla Q_k(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i})$ is amplification of the gradient noise term $\mu\, \boldsymbol{s}_{k,i}$ by an additive term $\boldsymbol{q}_{k,i}$.

1.1. Related Works

Solutions to the aggregate optimization problem (2) can be pursued by a variety of decentralized algorithms, including primal [1–3, 8] and primal-dual [9–13] methods.

The notion of ϵ-differential privacy as a means of quantifying the privacy loss encountered by sharing functions of private data is due to [7, 14], as is the Laplace mechanism, which ensures ϵ-differential privacy by perturbing the output of the function by Laplacian noise, where the power of the perturbation is calibrated to the sensitivity of the function and the desired privacy level ϵ.

In the context of centralized optimization by means of recursive algorithms, differential privacy has been applied to gradient descent [15–17], deep learning [18], as well as federated learning [19, 20]. The decentralized setting considered in this work is studied in [6, 21–24], where independent and identically distributed perturbations are added at each agent as in (8) and differential privacy is established.

Similarly to these related works, our scheme is based on stochastic gradient descent, and employs perturbations to achieve privacy. In contrast to prior works, however, locally generated perturbations at each agent will be tuned to the local graph topology, ensuring that the effect on the evolution of the network centroid is minimized, while preserving privacy guarantees. The authors in [25] present a "topology-aware" perturbation scheme, where noise powers are tuned to the local connectivity of agents. We, on the other hand, will be constructing the actual realizations, rather than perturbation powers, to match the graph topology.

2. DIFFUSION WITH GRAPH-HOMOMORPHIC PERTURBATIONS

We generalize the scheme (11)–(13), and allow agent ℓ to send different perturbation vectors $\boldsymbol{q}_{\ell k,i}$ to different neighbors k, resulting in:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \nabla J_k(\boldsymbol{w}_{k,i-1}) - \mu\, \boldsymbol{s}_{k,i} \tag{14}$$

$$\boldsymbol{\psi}_{k\ell,i} = \boldsymbol{\phi}_{k,i} + \boldsymbol{q}_{k\ell,i} \tag{15}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\psi}_{\ell k,i} \tag{16}$$

Our objective is to exploit this additional degree of freedom to construct the perturbations $\boldsymbol{q}_{\ell k,i}$ in a manner that protects agent ℓ from agent k, but minimizes the negative effect on the network as a whole. Previous studies on the dynamics of the diffusion recursion without privacy guarantees have shown that the local dynamics of each agent closely track that of a network centroid after sufficient iterations, both in the convex [2,3] and nonconvex [4,5] setting. From (16), we find for the network centroid:

$$\begin{aligned}
\boldsymbol{w}_{c,i} &\triangleq \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{w}_{k,i} \\
&\overset{(16)}{=} \frac{1}{K}\sum_{k=1}^{K}\sum_{\ell=1}^{K} a_{\ell k}\boldsymbol{\phi}_{\ell,i} + \frac{1}{K}\sum_{k=1}^{K}\sum_{\ell=1}^{K} a_{\ell k}\boldsymbol{q}_{\ell k,i} \\
&= \frac{1}{K}\sum_{\ell=1}^{K}\left(\sum_{k=1}^{K} a_{\ell k}\right)\boldsymbol{\phi}_{\ell,i} + \frac{1}{K}\sum_{\ell=1}^{K}\sum_{k=1}^{K} a_{\ell k}\boldsymbol{q}_{\ell k,i} \\
&\overset{(5)}{=} \frac{1}{K}\sum_{\ell=1}^{K}\boldsymbol{\phi}_{\ell,i} + \frac{1}{K}\sum_{\ell=1}^{K}\sum_{k=1}^{K} a_{\ell k}\boldsymbol{q}_{\ell k,i} \\
&\overset{(7)}{=} \boldsymbol{w}_{c,i-1} - \frac{\mu}{K}\sum_{\ell=1}^{K}\nabla Q(\boldsymbol{w}_{k,i-1};\boldsymbol{x}_{k,i}) \\
&\quad + \frac{1}{K}\sum_{\ell=1}^{K}\sum_{k=1}^{K} a_{\ell k}\boldsymbol{q}_{\ell k,i}
\end{aligned} \tag{17}$$

We observe that the network centroid $\boldsymbol{w}_{c,i}$ evolves similarly to a stochastic gradient update on the aggregate loss (2), perturbed by the sample mean of the weighted privacy terms $a_{\ell k}\boldsymbol{q}_{\ell k,i}$. The key question then is whether it is possible to construct $\boldsymbol{q}_{\ell k,i}$ in an uncoordinated manner, such that:

$$\frac{1}{K}\sum_{\ell=1}^{K}\sum_{k=1}^{K} a_{\ell k}\boldsymbol{q}_{\ell k,i} \overset{\text{desired}}{=} 0 \tag{18}$$

while preserving the privacy of all agents. If this were the case, the evolution of the network centroid would be largely unaffected by the privacy perturbations. We say "largely unaffected" since the gradients $\nabla Q(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i})$ are evaluated at $\boldsymbol{w}_{k,i-1}$, rather than $\boldsymbol{w}_{c,i-1}$ and hence indirectly affected by the privacy perturbations. As such, a more detailed performance analysis is necessary, which we conduct further below. The key take-away from the analysis will be that, despite the fact that the perturbations added in (15) are independent of the step-size, ensuring (18) modulates the effect of the privacy perturbations on the evolution of the centroid by a factor of the step-size μ, allowing for increasing levels of privacy perturbation as the step-size decreases. Since perturbations satisfying (18) have a reduced effect on the evolution of the network centroid under the combination matrix A, we will refer to them as "graph-homomorphic".

Definition 1 (Graph-homomorphic perturbations). *A set of perturbations $\boldsymbol{q}_{\ell k,i}$ is homomorphic for the the graph de-*

fined by the combination matrix $A \triangleq [a_{\ell k}]$ if it holds with probability one that:

$$\frac{1}{K}\sum_{\ell=1}^{K}\sum_{k=1}^{K} a_{\ell k}\boldsymbol{q}_{\ell k,i} = 0 \tag{19}$$

While other constructions are possible, we present here a simple construction, which can be implemented locally and independently at every agent k.

Lemma 1 (Constructing graph-homomorphic perturbations). *Let each agent ℓ sample independently from the Laplace distribution $\boldsymbol{v}_{\ell,i} \sim \mathrm{Lap}\,(0, b_v)$ with variance $\sigma_v^2 = 2b_v^2$. Then, the construction:*

$$\boldsymbol{q}_{\ell k,i} = \begin{cases} \boldsymbol{v}_{\ell,i}, & \text{if } k \in \mathcal{N}_\ell \text{ and } k \neq \ell, \\ -\frac{1-a_{\ell\ell}}{a_{\ell\ell}}\boldsymbol{v}_{\ell,i}, & \text{if } k = \ell. \end{cases} \tag{20}$$

is homomorphic for the graph described by the symmetric adjacency matrix $A = A^{\mathsf{T}}$.

Proof. The result can be verified immediately by substitution. □

3. ANALYSIS

3.1. Modeling Conditions

We make the following common assumptions to facilitate the performance and privacy analysis.

Assumption 1 (Adjacency matrix). *The combination matrix $A \triangleq [a_{\ell k}]$ is symmetric and doubly-stochastic, i.e.:*

$$a_{\ell k} = a_{k\ell},\ \sum_{\ell\in\mathcal{N}_k} a_{\ell k} = 1,\ a_{\ell k} = 0\ \forall\ \ell \notin \mathcal{N}_k \tag{21}$$

Furthermore, the graph described by A is connected, ensuring that:

$$\lambda_2 \triangleq \rho\left(A - \frac{1}{K}\mathbb{1}\mathbb{1}^{\mathsf{T}}\right) < 1 \tag{22}$$

Assumption 2 (Smoothness). *The risk functions $Q(\cdot;\boldsymbol{x}_k)$ have uniformly Lipschitz gradients, i.e., for all w_1, w_2, and with probability one:*

$$\|\nabla Q(w_1;\boldsymbol{x}_k) - \nabla Q(w_2;\boldsymbol{x}_k)\| \leq \delta\|w_1 - w_2\| \tag{23}$$

Additionally, we impose a bound on the norm of the gradient:

$$\|\nabla Q(w;\boldsymbol{x}_k)\| \leq G \tag{24}$$

3.2. Privacy Analysis

We now quantify the privacy loss encountered by a particular agent, when deciding to participate in the learning protocol. To quantify privacy precisely, we will employ the notion of ϵ-differential privacy [7]. For simplicity of exposition, and without loss of generality, we will focus on establishing a privacy guarantee for agent 1. By symmetry, the same argument applies to all other agents as well.

To this end, consider an alternative scenario, where agent 1 has decided not to volunteer its private information for the diffusion of information, and its data $\boldsymbol{x}_1$ is replaced by some other data $\boldsymbol{x}_1'$, following a different distribution. In this setting, implementing (14)–(16), would naturally result in a different learning trajectory $\boldsymbol{w}_{k,i}'$ at every agent k, since the data $\boldsymbol{x}_1'$ propagates through $\nabla Q(\boldsymbol{w}_{1,i}', \boldsymbol{x}_1')$ and the diffusion of estimates through the entire network. We first quantify the sensitivity of the evolution of the algorithm (14)–(16), a quantity that determines the amount of perturbation necessary to mask any particular agent [6, 7].

Lemma 2 (Sensitivity of the diffusion algorithm). *The distance between the trajectories $\boldsymbol{w}_{k,i}$ and $\boldsymbol{w}_{k,i}'$ is bounded with probability one by:*

$$\Delta(i) \triangleq \max_k \|\,\boldsymbol{w}_{k,i} - \boldsymbol{w}_{k,i}'\,\| \leq \mu 2Gi \tag{25}$$

Proof. Omitted due to space limitations. □

Definition 2 (ϵ-differential privacy). *We say that the diffusion recursion (14)–(16) is $\epsilon(i)$-differentially private for agent 1 at time i if:*

$$\frac{f\left(\left\{\left\{\boldsymbol{\psi}_{1\ell,n}\right\}_{\ell\neq 1\in\mathcal{N}_1}\right\}_{n=0}^{i}\right)}{f\left(\left\{\left\{\boldsymbol{\psi}'_{1\ell,n}\right\}_{\ell\neq 1\in\mathcal{N}_1}\right\}_{n=0}^{i}\right)} \leq e^{\epsilon(i)} \tag{26}$$

where $f(\cdot)$ denotes the probability density function and $\left\{\left\{\boldsymbol{\psi}_{1\ell,n}\right\}_{\ell\neq 1\in\mathcal{N}_1}\right\}_{n=0}^{i}$ collects all quantities transmitted by agent 1 to any of its neighbors during the operation of the algorithm, while excluding its local iterates $\boldsymbol{\psi}_{11,n}$, which are kept private.

In light of the fact that $e^{\epsilon(i)} \approx 1 - \epsilon(i)$ for small $\epsilon(i)$, relation (26) ensures that the distribution of estimates shared by agent 1 is close to unaffected (for small $\epsilon(i)$), whether agent 1 uses its own private data $\boldsymbol{x}_1$, or a proxy $\boldsymbol{x}_1'$, and as such little can be inferred about $\boldsymbol{x}_1$ by observing messages shared by agent 1.

Theorem 1 (Privacy cost of the diffusion algorithm). *Suppose (14)–(16) employs homomorphic perturbations constructed as in (20). Then, at time i, algorithm (14)–(16) is $\epsilon(i)$-differentially private according to (26), with:*

$$\epsilon(i) = \mu\frac{G(i^2+i)}{b_v} \tag{27}$$

Proof. Omitted due to space limitations. □

3.3. Performance Analysis

In order to quantify the impact of the privacy perturbations on the performance of the algorithm, we conduct a performance analysis in the presence of perturbations. Following the arguments in [4] for analyzing the dynamics of the unperturbed recursion (3)–(4) in nonconvex environments, we begin by establishing that the collection of iterates $\{\boldsymbol{w}_{k,i}\}_{k=1}^{K}$ continue to cluster around the network centroid $\boldsymbol{w}_{c,i}$.

Lemma 3 (Network disagreement). *Suppose the collection of agents $\{\boldsymbol{w}_{k,i}\}_{k=1}^{K}$ is initialized at a common, non-informative location, say $\boldsymbol{w}_{k,0} = \operatorname{col}\{0,\ldots,0\}$ for all k. Then, the deviation from the centroid is bounded for all $i \geq 0$ as:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\,\|\boldsymbol{w}_{k,i}-\boldsymbol{w}_{c,i}\|^2 \leq \mu^2\frac{\lambda_2^2}{(1-\lambda_2)^2}G^2 + b_v^2\frac{2\overline{a}}{1-\lambda_2} \tag{28}$$

where:

$$\overline{a} \triangleq \max_k\left\{(1-a_{kk}) + \frac{(1-a_{kk})^2}{a_{kk}^2}\right\} \tag{29}$$

Proof. Omitted due to space limitations. □

Relative to performance expressions for non-private decentralized gradient descent, we observe that the privacy perturbations account for an additional deviation on the order of $O(b_v^2)$. Nevertheless, relation (18) under (20) allows us to establish an improved descent relation.

Theorem 2 (Descent relation). *Under Assumptions 1–2, and for homomorphic perturbations constructed as in* (20)*, the network centroid descends along the loss* (2) *as:*

$$\begin{aligned}\mathbb{E}\,J(\boldsymbol{w}_{c,i}) \leq\ & \mathbb{E}\,J(\boldsymbol{w}_{c,i-1}) - \frac{\mu}{2}(1-2\mu\delta)\,\mathbb{E}\,\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2 \\ & + \frac{\mu}{2}(1+2\delta\mu)b_v^2\frac{2\delta^2\overline{a}}{1-\lambda_2} + \mu^2 2\delta G^2 + O(\mu^3)\end{aligned} \tag{30}$$

Proof. Omitted due to space limitations. □

Examination of (30) reveals that, despite the fact that the amount of perturbations added in (15) is independent of the step-size, their negative effect on the ability of the network centroid to descend along the aggregate loss $J(w)$ is multiplied by μ, and hence decays with the step-size.

Corollary 1 (Convergence to stationary points). *Suppose $J(w) \geq J^o$. Then, under Assumptions 1–2, and for homomorphic perturbations constructed as in* (20)*, we have:*

$$\frac{1}{i}\sum_{n=0}^{i-1}\mathbb{E}\,\|\nabla J(\boldsymbol{w}_{c,n}\,\|^2 \leq O\left(\frac{1}{\mu i}\right) + O(b_v^2) + O(\mu G^2) \tag{31}$$

Proof. The result follows after rearranging (30) and telescoping. □

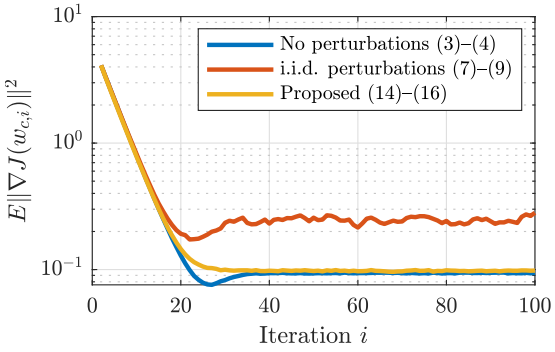

Fig. 1: Performance comparison with $M = 5$, $K = 20$, $\mu = 1$, $\rho = 0.1$, $\sigma_h^2 = 1$, $\sigma_p^2 = 2$.

4. NUMERICAL RESULTS

We verify the analytical results in the context of decentralized logistic regression for binary classification. Given class labels $\boldsymbol{\gamma} \in \{+1,-1\}$, we construct feature vectors $\boldsymbol{h}$ to be conditionally Gaussian, with means μ_{+1} and μ_{-1} respectively, i.e., $\boldsymbol{h} \in \mathbb{R}^M$ with $f(\boldsymbol{h}|\boldsymbol{\gamma}=\gamma) = \mathcal{N}\left(\mu_\gamma, \sigma_h^2\right)$. Each agent k is equipped with a local logistic loss function of the form:

$$J_k(w) \triangleq \mathbb{E}\ln\left(1+e^{-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}}w}\right) + \frac{\rho}{2}\|w\|^2 \tag{32}$$

We compare the performance of the ordinary diffusion recursion (3)–(4) with the privatized recursion (7)–(9) and the proposed scheme (14)–(16), constructed according to (20). The resulting performance is illustrated in Fig. 1. We observe that the proposed perturbation scheme approximately matches the performance of the non-private diffusion implementation, while outperforming the implementation with i.i.d. perturbations, despite employing the same perturbation powers σ_p^2.

5. CONCLUSION

We have proposed a new perturbation scheme for differentially private decentralized stochastic optimization, where the perturbations are constructed at each agent to match the local graph topology. The resulting perturbations are invisible to the network centroid under the diffusion operation, while preserving ϵ-differential privacy, and hence termed graph-homomorphic (for a particular topology). Analytical and numerical results show that the construction reduces the negative effect of privacy perturbations, while preserving differential privacy.

6. REFERENCES

[1] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, July 2014.

[2] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks - Part I: Transient analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3487–3517, June 2015.

[3] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks – Part II: Performance analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3518–3548, June 2015.

[4] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments – Part I: Agreement at a Linear rate," *to appear in IEEE Transactions on Signal Processing, available as arXiv:1907.01848*, 2021.

[5] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments – Part II: Polynomial escape from saddle-points," *to appear in IEEE Transactions on Signal Processing, available as arXiv:1907.01849*, 2021.

[6] Z. Huang, S. Mitra, and N. Vaidya, "Differentially private distributed optimization," in *Proc. International Conference on Distributed Computing and Networking*, Goa, India, Jan. 2015, pp. 1–10.

[7] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, Aug. 2014.

[8] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan 2009.

[9] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[10] P. Di Lorenzo and G. Scutari, "Next: In-network non-convex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.

[11] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning – Part II: Convergence analysis," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 724–739, Feb 2019.

[12] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *Proc. IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 8353–8358.

[13] D. Jakovetić, D. Bajović, J. Xavier, and J. M. F. Moura, "Primal-dual methods for large-scale and distributed convex optimization and data analytics," *Proceedings of the IEEE*, pp. 1–16, 2020.

[14] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, Berlin, Heidelberg, 2006, pp. 265–284, Springer Berlin Heidelberg.

[15] A. Rajkumar and S. Agarwal, "A differentially private stochastic gradient descent algorithm for multiparty classification," in *Proc. Machine Learning Research*, La Palma, Canary Islands, Apr 2012, pp. 933–941.

[16] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *Proc. IEEE Global Conference on Signal and Information Processing*, 2013, pp. 245–248.

[17] J. Lee and D. Kifer, "Concentrated differentially private gradient descent with adaptive per-iteration privacy budget," in *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2018, pp. 1656–1665.

[18] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, 2016, p. 308–318.

[19] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *available as arXiv:1712.07557*, Dec 2017.

[20] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[21] T. Zhang and Q. Zhu, "Dynamic differential privacy for admm-based distributed classification learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 172–187, 2017.

[22] C. Li, P. Zhou, L. Xiong, Q. Wang, and T. Wang, "Differentially private distributed online learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 8, pp. 1440–1453, 2018.

[23] M. Showkatbakhsh, C. Karakus, and S. Diggavi, "Differentially private consensus-based distributed optimization," *available as arXiv:1903.07792*, March 2019.

[24] M. Hou, D. Li, X. Wu, and X. Shen, "Differential privacy of online distributed optimization under adversarial nodes," in *Proc. Chinese Control Conference (CCC)*, 2019, pp. 2172–2177.

[25] T. Xiang Y. Liu S. Guo, T. Zhang, "Differentially private decentralized learning," *available as arXiv:2006.07817*, June 2020.