DISTRIBUTED NONCONVEX LEARNING UNDER COUPLING CONSTRAINTS OVER DIRECTED NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributed learning is often formulated through consensus, yet many collaborative models are built from distinct local components that must satisfy global consistency relations rather than agree with one another. This becomes particularly challenging when the objectives are nonconvex, gradients are stochastic, and communication is directed and unbalanced, since global feasibility evolves with the local models and cannot be evaluated at any single node. We study this setting through general affine coupling and propose DiSPART (Directed Stochastic Primal–dual Algorithm with Residual Tracking). DiSPART turns global constraint information into distributed feedback by combining push–sum residual and multiplier tracking with gradient smoothing and moving-center regularization. Under standard smoothness, constraint regularity, bounded-variance stochastic gradients, and an explicit Perron-weighted mixing condition, we establish a finite-time bound on the average expected squared KKT residual that jointly controls stationarity and feasibility. For fixed problem and network parameters, the bound gives $\mathcal{O}(K^{-1/2})$ convergence with stochastic gradients and $\mathcal{O}(K^{-1})$ with exact gradients, while allowing infeasible primal initialization and a horizon-independent penalty without assuming bounded multipliers. The analysis further makes explicit how affine coupling, network mixing, and Perron imbalance shape finite-time performance. Experiments support the algorithmic design and the predicted network effects, while two real-data classification tasks show that these guarantees translate into practical constrained learning with substantially lower gradient-query costs than exact-gradient baselines.

1 INTRODUCTION

Distributed learning enables multiple participants to train models through local computation and peer-to-peer communication (Assran et al., 2019; Wang et al., 2025). For modern neural models, the resulting objectives are typically nonconvex, while full local gradients can be expensive to evaluate. Stochastic mini-batch updates provide a natural alternative, but introduce gradient noise that must be controlled throughout training (Sun et al., 2020; Xin et al., 2021). Decentralized schemes further avoid repeated global aggregation, yet their standard consensus formulation can be restrictive when participants optimize different model components.

Many collaborative-learning tasks instead require compatibility among heterogeneous local variables. Vertical federated learning provides a representative example, where participants hold distinct feature blocks for aligned samples and jointly construct a prediction without requiring their local parameters to agree (Valdeira et al., 2025). For an additive predictor, introducing a shared variable ζ yields the affine consistency relation $\sum_i F_i w_i - \zeta = 0$ (Yarmoshik et al., 2025). This relation remains affine when ζ serves as a linear preactivation for a smooth nonlinear prediction layer, motivating general affine coupling as a natural model beyond consensus.

Without a central coordinator, such coupling must be maintained over the communication network itself. Directed and asymmetric interactions motivate column-stochastic mixing, but nonuniform Perron weights bias uncorrected local estimates, while slow mixing delays the propagation of global information (Assran et al., 2019; Liang et al., 2025). Push–sum corrects the weighting bias, yet the global feasibility residual still evolves with the stochastic local updates and must be tracked across the network. The resulting challenge is to achieve stationarity and global feasibility simultaneously

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

under nonconvexity, stochastic gradients, affine coupling, and directed unbalanced communication. We study the following distributed nonconvex learning problem:

$$\min_{x_1,\dots,x_n} f(x) := \sum_{i=1}^{n} f_i(x_i), \qquad \text{s.t.} \quad \sum_{i=1}^{n} A_i x_i = b. \tag{1}$$

where $f_i(x_i) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[\phi_i(x_i;\xi_i)]$ is the local expected objective, with $\xi_i \sim \mathcal{D}_i$ denoting a local data sample and $\phi_i(\cdot;\xi_i)$ the corresponding sample loss. Each f_i is assumed to be smooth, with no convexity assumption imposed. The variable $x_i \in \mathbb{R}^{d_i}$ denotes the local decision variable, while $A_i \in \mathbb{R}^{m\times d_i}$ specifies its contribution to the global affine constraint, with $b \in \mathbb{R}^m$.

Table 1 compares representative guarantees across the main dimensions of our setting. SGP and MG-Push-DIGing address stochastic nonconvex optimization of a common model over directed networks (Assran et al., 2019; Liang et al., 2025). For coupled problems, Yarmoshik et al. (2025) study general affine block coupling under strong convexity, while D-SMPL/D-SCAMPL consider stochastic nonconvex functional constraints on a shared variable over undirected networks (Sharma et al., 2026). DiSPART targets stochastic nonconvex learning with general affine block coupling over unbalanced directed communication, with finite-time KKT guarantees.

Table 1: **Representative methods: settings and convergence guarantees.**

Method	NC	SG	GC	UN	Criterion	Leading Rate
SGP (Assran et al., 2019)	✓	✓	–	✓	Average squared gradient norm	$\mathcal{O}(K^{-1/2})$
MG-Push-DIGing (Liang et al., 2025)	✓	✓	–	✓	Average squared gradient norm	$\widetilde{\mathcal{O}}(K^{-1/2}+K^{-1})$
Yarmoshik et al. (2025)	–	–	✓	–	Squared primal distance	Linear (strongly convex)
D-SMPL / D-SCAMPL (Sharma et al., 2026)	✓	✓	–	–	Approximate KKT	$\mathcal{O}(K^{-2/3})$
DiSPART	✓	✓	✓	✓	Average squared KKT residual	$\mathcal{O}(K^{-1/2})+\mathcal{O}(K^{-1})^{*}$

Notes. NC: nonconvex objectives; SG: stochastic gradients; GC: affine coupling across distinct local blocks (excluding consensus); UN: unbalanced directed networks. Rates are stated under the respective criteria and assumptions of each source; stochastic rates are in expectation, and per-iteration costs may differ. The notation $\widetilde{\mathcal{O}}$ suppresses logarithmic factors. *Theorem 1 uses horizon-dependent parameters that remain fixed within each run.

Beyond this setting-level comparison, the stabilization used in DiSPART is motivated by smoothed proximal primal–dual methods for nonconvex constrained optimization, including S-prox-ALM (Zhang & Luo, 2022) and related stochastic variants (Huang et al., 2025). Here, this mechanism is coupled with stochastic gradient estimation and distributed affine-residual tracking over a column-stochastic network.

Contributions. Our main contributions are as follows:

- **Algorithm design.** We develop DiSPART for stochastic nonconvex learning with heterogeneous local variables under general affine coupling over directed column-stochastic networks, extending decentralized learning beyond the standard consensus setting.
- **Convergence guarantee.** We establish finite-time convergence guarantees in terms of the KKT residual, without assuming bounded multipliers or feasible primal initialization. The analysis explicitly quantifies the effects of affine coupling, network mixing, and Perron imbalance on convergence.
- **Experimental validation.** We validate DiSPART through component ablations, comparisons with existing methods, and controlled network studies, supporting the proposed design and highlighting the distinct effects of network mixing and Perron imbalance.

2 PROBLEM SETUP

Notation. Let $x := \mathrm{col}(x_1, \dots, x_n) \in \mathbb{R}^d$ denote the stacked decision vector, where $d := \sum_{i=1}^n d_i$, and let $A := [A_1, \dots, A_n] \in \mathbb{R}^{m \times d}$ denote the coupling matrix. Define the constraint residual as $r(x) := Ax - b$, and set $A_0 := \|A\|$, $A_{\max} := \max_i \|A_i\|$, and $\sigma_A := \sigma_{\min}(A)$. Any fixed decomposition $b = \sum_{i=1}^n b_i$ may be used, with $b_i \in \mathbb{R}^m$ and both A_i and b_i locally available at node i. Throughout, $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced spectral norm for matrices. We use $\mathbf{1}$ for the all-ones vector, I_p for the $p \times p$ identity matrix, $\mathrm{diag}(\cdot)$ for the diagonal-matrix operator, and $\otimes$ for the Kronecker product.

Communication topology. The communication network is represented by a fixed, strongly connected directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $(j, i) \in \mathcal{E}$ indicates that node j can transmit information to node i. Each node j assigns nonnegative weights w_{ij} to its outgoing information, with $w_{ij} = 0$ whenever $i \neq j$ and $(j, i) \notin \mathcal{E}$. These weights are normalized at the sending node so that $\sum_{i=1}^n w_{ij} = 1$. The resulting mixing matrix $W = [w_{ij}] \in \mathbb{R}^{n \times n}$ is therefore column-stochastic, satisfying $\mathbf{1}^\top W = \mathbf{1}^\top$. Let $\pi \in \mathbb{R}^n_{++}$ denote the normalized right Perron vector of W, satisfying $W\pi = \pi$ and $\mathbf{1}^\top \pi = 1$, and set $\Pi := \mathrm{diag}(\pi)$. Following Liang et al. (2025), we characterize the network through the equilibrium skewness κ_π and the generalized spectral gap $1 - \beta_\pi$, where

$$\kappa_\pi := \frac{\pi_{\max}}{\pi_{\min}}, \qquad \beta_\pi := \left\| \Pi^{-1/2} \big(W - \pi \mathbf{1}^\top \big) \Pi^{1/2} \right\|_2,$$

with $\pi_{\max} := \max_i \pi_i$ and $\pi_{\min} := \min_i \pi_i$. The equilibrium skewness measures the nonuniformity of the Perron weights, while the generalized spectral gap describes the mixing strength of W. When W is doubly stochastic, $\pi = \mathbf{1}/n$, so that $\kappa_\pi = 1$ and $\beta_\pi = \|W - \mathbf{1}\mathbf{1}^\top/n\|_2$.

Assumption 1 (Smoothness). *Each $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ is L_i-smooth on $\mathbb{R}^{d_i}$; that is,*

$$\|\nabla f_i(u) - \nabla f_i(v)\| \leq L_i \|u - v\|, \qquad \forall\, u, v \in \mathbb{R}^{d_i}.$$

This assumption bounds the variation of each local gradient over its entire domain without requiring convexity. We write $L := \max_i L_i$ for the common smoothness constant.

Assumption 2 (Constraint regularity). *The coupling matrix $A \in \mathbb{R}^{m \times d}$ satisfies $\mathrm{rank}(A) = m$, and the constrained optimal value $f_* := \inf_{Ax=b} f(x)$ is finite.*

The full-row-rank condition ensures feasibility. Lower boundedness of the objective is required only on the feasible set.

Assumption 3 (Network mixing). *The mixing matrix W is fixed, nonnegative, and column-stochastic, with a normalized positive right Perron vector π. Furthermore,*

$$\beta_\pi < 1.$$

This condition ensures that each communication step reduces the deviation from the network equilibrium in the Perron-weighted norm. A smaller β_π indicates stronger mixing.

Let $\mathcal{F}_k$ denote the σ-algebra representing the information available before sampling ξ_i^{k+1}, and let x_i^{k+1} be an $\mathcal{F}_k$-measurable query point.

Assumption 4 (Stochastic oracle). *For each i, there exists a constant $\sigma_i \geq 0$ such that, for every $k \geq 0$,*

$$\begin{aligned} &\mathbb{E}\big[\nabla \phi_i(x_i^{k+1}; \xi_i^{k+1}) \mid \mathcal{F}_k\big] = \nabla f_i(x_i^{k+1}), \\ &\mathbb{E}\big[\|\nabla \phi_i(x_i^{k+1}; \xi_i^{k+1}) - \nabla f_i(x_i^{k+1})\|^2 \mid \mathcal{F}_k\big] \leq \sigma_i^2. \end{aligned} \tag{2}$$

The same conditions hold initially, conditional on the information before sampling.

The stochastic gradients are conditionally unbiased with uniformly bounded noise variance. We write $\sigma^2 := \sum_i \sigma_i^2$ for the aggregate variance bound.

3 THE DISPART ALGORITHM

This section introduces the Directed Stochastic Primal–dual Algorithm with Residual Tracking (DiSPART), which is specifically developed to address the coupled challenges of stochastic nonconvex optimization, affine equality constraints, and unbalanced directed communication. Within this setting, each agent has access only to local information, whereas the primal and dual updates rely on global constraint residuals that cannot be evaluated locally. To stabilize the resulting stochastic nonconvex primal–dual dynamics, DiSPART employs two complementary mechanisms: gradient momentum mitigates the influence of sampling fluctuations, while a proximal regularization term around a moving center enhances robustness to nonconvex curvature. By integrating distributed residual tracking, stochastic variance control, and primal–dual stabilization into a unified framework, DiSPART offers a scalable approach for constrained optimization over directed networks. The complete procedure is summarized in Algorithm 1.

Algorithm 1 DiSPART

1: At each node i, choose deterministic $x_i^0 \in \mathbb{R}^{d_i}$ and initialize $z_i^0 = x_i^0$ and $v_i^0 = 1$.
2: Draw ξ_i^0 and set $g_i^0 = \nabla \phi_i(x_i^0; \xi_i^0)$.
3: Set $s_i^0 = n(A_i x_i^0 - b_i)$, $\ell_i^0 = 0$, $q_i^0 = s_i^0 / v_i^0$ and $\lambda_i^0 = \ell_i^0 / v_i^0$.
4: Choose $\alpha, \tau, \gamma, \rho > 0$, $0 < \beta \leq 1$, and $\omega > L$.
5: **for** $k = 0, \ldots, K-1$ **do**
6: Update the local primal blocks:
$$x_i^{k+1} = x_i^k - \alpha\big[g_i^k + A_i^\top(\lambda_i^k + \rho q_i^k) + \omega(x_i^k - z_i^k)\big].$$
7: Compute the local residual increments:
$$\Delta h_i^k = A_i(x_i^{k+1} - x_i^k).$$
8: Exchange the following messages:
$$v_i^k,\ s_i^k + n\Delta h_i^k,\ \ell_i^k.$$
9: Update the push–sum weights and residual numerators, and normalize:
$$v_i^{k+1} = \sum\nolimits_{j=1}^n w_{ij} v_j^k, \quad s_i^{k+1} = \sum\nolimits_{j=1}^n w_{ij}(s_j^k + n\Delta h_j^k), \quad q_i^{k+1} = s_i^{k+1}/v_i^{k+1}.$$
10: Update the multiplier numerators and normalize:
$$\ell_i^{k+1} = \sum\nolimits_{j=1}^n w_{ij}\ell_j^k + \tau s_i^{k+1}, \quad \lambda_i^{k+1} = \ell_i^{k+1}/v_i^{k+1}.$$
11: Update the proximal centers:
$$z_i^{k+1} = z_i^k + \gamma(x_i^{k+1} - z_i^k).$$
12: Draw a fresh sample ξ_i^{k+1} and update the gradient estimates:
$$g_i^{k+1} = (1-\beta)g_i^k + \beta\nabla\phi_i(x_i^{k+1}; \xi_i^{k+1}).$$
13: **end for**

DiSPART coordinates local primal–dual updates through distributed estimates of the global constraint residual. Since each node can evaluate only its own contribution to this residual, the tracking recursion in Algorithm 1, lines 7–9, incorporates local residual increments to maintain these estimates as the primal variables evolve. Push–sum normalization compensates for the imbalance of column-stochastic communication, and the multiplier update in line 10 reuses the updated residual numerator. Both recursions thus share a single communication exchange, without requiring any node to evaluate the global residual explicitly.

This distributed constraint feedback enters a primal step regularized around a moving proximal center (lines 6 and 11). The proximal correction counteracts nonconvex curvature, while the evolving center avoids anchoring the iterates to a prescribed reference point. At a joint fixed point, the center coincides with the primal variable and the correction vanishes. The regularization therefore supports the primal dynamics without introducing an additional term into the stationarity conditions of the original problem. Alongside this curvature control, the momentum-based gradient update in line 12 addresses the fluctuations introduced by stochastic sampling. It combines past gradient information with a fresh sample at the updated primal iterate, moderating sampling noise with only one new gradient sample per node and iteration. Since the sampling points change with the primal variables, the

estimator must also follow the evolving gradient. Its weighting therefore balances noise attenuation against the lag in tracking the current gradient.

Together, these updates form a coupled feedback loop: primal changes drive residual tracking and gradient estimation, whose errors in turn influence subsequent primal–dual steps. The convergence analysis accounts for this interaction together with the moving-center dynamics to establish guarantees under suitable parameter choices.

For the convergence analysis that follows, the nodewise updates are collected into a stacked recursion. For $u \in \{x, g, z, v, s, \ell, q, \lambda\}$, let $u^k := \mathrm{col}_{i=1}^n(u_i^k)$ denote the corresponding stacked state. The local constraint mappings and network mixing are represented by $\mathcal{A} := \mathrm{blkdiag}(A_1, \dots, A_n)$ and $W_m := W \otimes I_m$, respectively, while the push–sum weights define the scaling matrix $\mathcal{V}^k := \mathrm{diag}(v^k) \otimes I_m$. With $\widehat{g}^{k+1} := \mathrm{col}_{i=1}^n\big(\nabla\phi_i(x_i^{k+1}; \xi_i^{k+1})\big)$ denoting the stacked fresh stochastic gradient, the DiSPART updates can be written, for $k = 0, \dots, K-1$, as

$$
\begin{aligned}
&\texttt{Primal update:} && x^{k+1} = x^k - \alpha\big[g^k + \mathcal{A}^\top(\lambda^k + \rho q^k) + \omega(x^k - z^k)\big],\\
&\texttt{Tracking:} && v^{k+1} = Wv^k, \quad s^{k+1} = W_m\big[s^k + n\mathcal{A}(x^{k+1} - x^k)\big],\\
&\texttt{Multiplier update:} && \ell^{k+1} = W_m\ell^k + \tau s^{k+1},\\
&\texttt{Normalization:} && q^{k+1} = (\mathcal{V}^{k+1})^{-1}s^{k+1}, \quad \lambda^{k+1} = (\mathcal{V}^{k+1})^{-1}\ell^{k+1},\\
&\texttt{Auxiliary updates:} && z^{k+1} = z^k + \gamma(x^{k+1} - z^k), \quad g^{k+1} = (1-\beta)g^k + \beta\widehat{g}^{k+1}.
\end{aligned}
$$

The residual recursion preserves aggregate consistency with the global constraint residual, while local updates remain subject to tracking and stochastic gradient-estimation errors. Gradient momentum smooths the stochastic feedback, while moving-center regularization compensates for nonconvex curvature. The following section analyzes these coupled dynamics and establishes convergence guarantees for DiSPART.

4 Convergence Guarantees

Convergence measure. For the nonconvex constrained problem, convergence is assessed through first-order stationarity and primal feasibility, consistent with KKT-based criteria for linearly constrained nonconvex optimization (Hong et al., 2017; Zhang & Luo, 2022). For fixed $\rho > 0$, define

$$
\bar{\lambda}^k := \frac{1}{n}(\mathbf{1}^\top \otimes I_m)\ell^k, \qquad \widehat{\lambda}^k := \bar{\lambda}^k + \rho r(x^k),
$$

where $\bar{\lambda}^k$ averages the multiplier numerators. The corrected multiplier $\widehat{\lambda}^k$ absorbs the penalty contribution into the stationarity condition of the original Lagrangian. Accordingly, define the squared KKT residual and its expected iteration average as

$$
R_k := \|\nabla f(x^k) + A^\top\widehat{\lambda}^k\|^2 + \|r(x^k)\|^2, \qquad M_K := \frac{1}{K}\sum\nolimits_{k=0}^{K-1}\mathbb{E}[R_k]. \tag{3}
$$

The quantity M_K averages the squared stationarity and feasibility errors over the first K iterations, with expectation taken over the stochastic oracle responses. Thus, a bound on M_K controls both errors in an averaged mean-square sense.

Coupling effects. Under affine coupling, distributed constraint enforcement must account for a global residual that evolves with the local primal variables. Primal increments induce local and global residual changes through A_i and A, respectively, while residual and multiplier tracking errors feed back into subsequent primal updates through $A_i^\top$. Controlling this interaction involves both the local block scale $A_{\max}$ and the aggregate scale A_0, which are combined in

$$
\Gamma_A := nA_{\max}^2 + \sqrt{n}\,A_{\max}A_0. \tag{4}
$$

For a fixed affine representation, Γ_A is independent of W. The following theorem quantifies the interaction of this coupling scale with network mixing and Perron imbalance through an explicit convergence bound.

Theorem 1 (**Convergence rate**). *Suppose Assumptions 1–4 hold, and let DiSPART be initialized as specified in Algorithm 1. For any fixed $\rho > 0$ and $\omega > L$, there exist positive coefficients $c_\tau, c_\gamma, c_\beta$*

depending only on ρ, ω and the fixed objective and coupling scales. With these coefficients fixed, a finite graph-dependent scale $H_\pi \geq 1$ can be chosen such that, for every prescribed horizon $K \geq 1$, running DiSPART with the constant parameters

$$\alpha = \alpha_K := \left(H_\pi + \sigma \sqrt{8 c_\beta K / D_0} \right)^{-1}, \quad \tau = c_\tau \alpha_K, \quad \gamma = c_\gamma \alpha_K, \quad \beta = c_\beta \alpha_K \tag{5}$$

yields

$$M_K \leq C \Bigg[\underbrace{\sigma \sqrt{\frac{D_0}{K}}}_{\text{sampling noise}} + \underbrace{\frac{D_0}{K}}_{\text{optimization error}} + \underbrace{\frac{D_0 (1 + \kappa_\pi \Gamma_A) \beta_\pi}{K (1 - \beta_\pi)}}_{\text{network penalty}} + \underbrace{\frac{C_{\mathrm{tr}}}{K} \left(1 + \frac{\kappa_\pi^2 \beta_\pi^2}{1 - \beta_\pi^2} \right)}_{\text{tracking transient}} \Bigg]. \tag{6}$$

Here, $D_0 > 0$ captures the initial optimization and gradient-estimation errors, while $C_{\mathrm{tr}} \geq 0$ represents the initial tracking mismatch. Both are independent of K and W, with definitions given in Appendix E.1. The constant $C > 0$ depends only on the fixed objective and coupling scales and on $\rho, \omega, c_\tau, c_\gamma, c_\beta$; it is independent of K, W, σ, and the initialization. Admissible coefficient choices and an explicit expression for C are provided in Appendix D.

Theorem 1 establishes an $\mathcal{O}(1/\sqrt{K})$ ergodic convergence rate for DiSPART in the average expected squared KKT residual, matching the iteration order of classical guarantees for smooth stochastic nonconvex optimization (Ghadimi & Lan, 2013). The *sampling noise* term reflects stochastic-gradient fluctuations, whose influence is balanced against optimization progress by the prescribed parameter scaling. When $\sigma = 0$, the oracle returns exact gradients and DiSPART becomes deterministic. The stochastic term vanishes, yielding an $\mathcal{O}(1/K)$ ergodic rate consistent in iteration order with deterministic linearly constrained nonconvex optimization results (Hong et al., 2017; Zhang & Luo, 2022). The *optimization error* term gives a graph-independent baseline determined by D_0, which accounts for the initial optimization and gradient-estimation errors, together with initial constraint violation. Greater constraint violation or larger initial error bounds enlarge the corresponding contributions to D_0 and hence the baseline bound. This contribution decreases at order $\mathcal{O}(1/K)$ in the iteration-averaged guarantee as the iteration budget grows.

The remaining two terms capture the additional effects of distributed constraint tracking over unbalanced directed networks. The *network penalty* reflects the cost of sufficient stepsize restrictions used to control feedback between the primal and tracking updates. A smaller generalized spectral gap $1 - \beta_\pi$ weakens the mixing guarantee and increases the bound through $\beta_\pi/(1 - \beta_\pi)$. Greater Perron imbalance and larger local and aggregate constraint scales further enlarge the bound through $\kappa_\pi \Gamma_A$. The *tracking transient*, by contrast, captures the accumulated influence of the initial tracking mismatch. Under the prescribed parameters, the cumulative squared contribution of this initial response is bounded independently of the horizon. Its factors κ_π^2 and $\beta_\pi^2/(1-\beta_\pi^2)$ arise from Perron-weighted norm comparisons and geometric-decay estimates, so greater imbalance or slower mixing can also enlarge the transient bound. Thus, network mixing and imbalance affect the coefficients of these two upper-bound contributions without changing their $\mathcal{O}(1/K)$ iteration order.

The following corollaries show how these contributions simplify under particular communication and coupling structures.

Corollary 1 (Doubly stochastic networks). *Under the conditions of Theorem 1, suppose that the matrix W is doubly stochastic. Then $\pi = \mathbf{1}/n$, $\kappa_\pi = 1$, and $v^k = \mathbf{1}$ throughout the run. The same parameter construction yields*

$$M_K \leq C \left[\sigma \sqrt{\frac{D_0}{K}} + \frac{D_0}{K} + \frac{D_0 (1 + n A_{\max}^2) \beta_\pi}{K (1 - \beta_\pi)} + \frac{C_{\mathrm{tr}}}{K (1 - \beta_\pi^2)} \right], \tag{7}$$

where $\beta_\pi = \|W - \mathbf{1}\mathbf{1}^\top/n\|_2$ and C, D_0, C_{tr} are as in Theorem 1.

Beyond eliminating Perron imbalance, doubly stochasticity removes the time-varying push–sum normalization, sharpening the explicit coupling coefficient in the network penalty from Γ_A to $n A_{\max}^2$. Mixing and initial tracking effects can nevertheless persist, showing that network balance alone does not eliminate the cost of distributed constraint tracking.

Corollary 2 (Centralization). *Under the conditions of Theorem 1, suppose that $n = 1$. Then the residual estimate is exact throughout the run, $\beta_\pi = 0$, and $C_{\mathrm{tr}} = 0$. Consequently,*

$$M_K \le C \left[\sigma \sqrt{\frac{D_0}{K}} + \frac{D_0}{K} \right]. \tag{8}$$

Compared with Corollary 1, centralization further eliminates the mixing and initial tracking contributions, since constraint feedback requires no communication. The coupling constraints remain, while (6) reduces to the sampling and initialization terms in (8), isolating the cost of distributed tracking. In the deterministic case, the resulting decay order of the squared KKT residual agrees with established guarantees for linearly constrained nonconvex optimization (Zhang & Luo, 2022).

Corollary 3 (Consensus constraints). *Suppose the assumptions and initialization of Theorem 1 hold, with $d_i = p$, $b_i = 0$ for all i, and $n \ge 2$. Let $A = C_{\mathrm{con}} \otimes I_p$, where $C_{\mathrm{con}} \in \mathbb{R}^{(n-1)\times n}$ has orthonormal rows spanning $\mathbf{1}^\perp$. Then $A_0 = \sigma_A = 1$ and $\Gamma_A = n - 1 + \sqrt{n-1}$. Retain the parameter ratios and the scale H_π from Theorem 1, and, for $\sigma > 0$, choose $\alpha = \alpha_K := \min\{H_\pi^{-1}, \sqrt{D_0/(8c_\beta \sigma^2 K)}\}$. With $f(u) := \sum_{i=1}^n f_i(u)$ and $\bar{x}^k := n^{-1}\sum_{i=1}^n x_i^k$,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(\bar{x}^k)\|^2 \le n C_{\mathrm{g}} \left[\sigma \sqrt{\frac{D_0}{K}} + \frac{D_0}{K} + \frac{D_0(1+\kappa_\pi \Gamma_A)\beta_\pi}{K(1-\beta_\pi)} + \frac{C_{\mathrm{tr}}}{K}\left(1 + \frac{\kappa_\pi^2 \beta_\pi^2}{1-\beta_\pi^2}\right)\right]. \tag{9}$$

Here $C_{\mathrm{g}} > 0$ is independent of K, W, σ, and the initialization. When $\sigma = 0$, take $\alpha_K = H_\pi^{-1}$; the same bound holds with the stochastic term equal to zero.

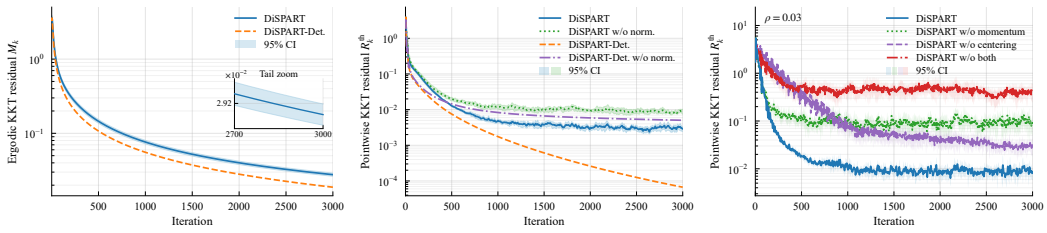
Under consensus constraints, the local variables are required to agree, making the mean iterate $\bar{x}^k$ a natural representative of the network state. Corollary 3 translates the general KKT guarantee into a stationarity bound for the aggregate objective evaluated at $\bar{x}^k$, which is a standard criterion in decentralized nonconvex optimization (Assran et al., 2019). Although disagreement does not appear as a separate term in (9), it is fully represented by the consensus residual under the orthonormal encoding and enters the conversion from local gradients to the gradient at the mean iterate through smoothness. The orthonormal encoding also removes arbitrary scaling of the consensus constraint by fixing $A_0 = \sigma_A = 1$, so that Γ_A depends only on the number of nodes. This makes the coupling contribution to the network penalty explicit, while the factors involving κ_π and β_π continue to reflect the influence of directed communication. For fixed n and network, (9) gives an $\mathcal{O}(K^{-1/2})$ bound on the time-averaged expected squared gradient norm at the mean iterate, improving to $\mathcal{O}(K^{-1})$ without oracle noise. Proofs of all three corollaries are provided in Appendix F.

5 Experiments

The experiments evaluate DiSPART on synthetic nonconvex problems with coupling constraints and two real-data classification tasks. The synthetic studies cover component ablations, method comparisons, and network effects, while the real-data experiments assess classification performance and constraint satisfaction. Full experimental details are provided in Appendix G.

Convergence and ablations. Figure 1 evaluates DiSPART on a ten-node nonconvex problem with coupling constraints. Panel (a) compares stochastic DiSPART with its exact-gradient counterpart, DiSPART-Det., using empirical estimates of M_K from Theorem 1. Panels (b,c) report the per-iteration KKT residual R_k to highlight late-iteration differences, with comparisons based on the mean over the final 200 iterations. Push–sum normalization reduces the residual by factors of 2.75 and 67.4 in the stochastic and deterministic settings, respectively (panel (b)). Under a separate fixed configuration, the complete method achieves 10.9-, 3.18-, and 43.2-fold reductions relative to variants without gradient momentum, moving-center regularization, and both, respectively (panel (c)). It also attains the lowest tail residual in all ten paired stochastic runs.

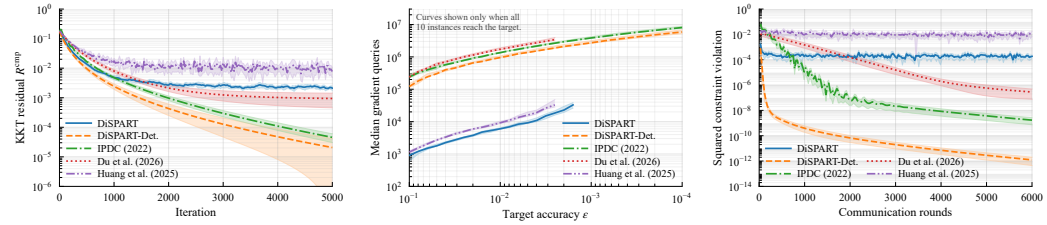
Comparison with existing methods. Figure 2 compares DiSPART with primal–dual methods for nonconvex problems with coupling constraints. IPDC (Zhang et al., 2022) and the method of Du et al. (2026) use exact gradients and are compared with DiSPART-Det., whereas Algorithm 1 of Huang et al. (2025) provides a stochastic counterpart to DiSPART. To provide a common optimality



(a) (b) (c)

Figure 1: **Convergence and component ablations of DiSPART.** (a) Empirical estimates of M_K, with an inset enlarging the stochastic curve. (b) Push–sum normalization ablation. (c) Gradient-momentum and moving-center ablations. Stochastic curves show means over ten oracle seeds with pointwise 95% Student-t confidence bands. Here, "w/o" denotes "without". Panels (a,b) share the complete-method trajectories, while panel (c) uses a separate configuration.

measure across methods, we use $R^{\mathrm{cmp}}(x) := \min_{\lambda \in \mathbb{R}^m} \|\nabla f(x) + A^\top \lambda\|^2 + \|r(x)\|^2$, consistent with the approximate KKT criterion in Zhang et al. (2022). As shown in panel (a), both DiSPART and DiSPART-Det. attain lower terminal mean residuals than their corresponding stochastic and deterministic baselines. Panel (b) further shows that stochastic DiSPART requires fewer median gradient queries to reach the common accuracy targets attained by all compared stochastic runs. Finally, at a matched communication budget, DiSPART achieves a lower final mean squared constraint violation than its stochastic comparator, while DiSPART-Det. attains the lowest mean constraint violation among all methods in panel (c).



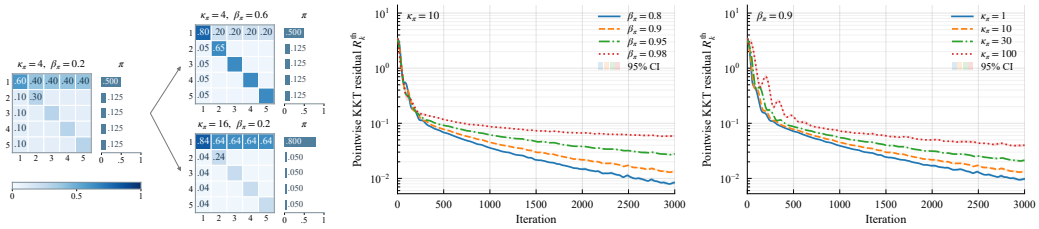
(a) (b) (c)

Figure 2: **Comparison with existing methods: convergence, query cost, and feasibility.** (a) Squared KKT residual versus outer iterations. (b) Gradient queries to target accuracy. (c) Squared constraint violation versus communication rounds. Results aggregate ten matched problem instances: panels (a,c) show means with pointwise 95% Student-t confidence intervals, while panel (b) shows median first-hit query costs with 95% percentile-bootstrap intervals.

Network effects. Figure 3 isolates the effects of network mixing and Perron imbalance by varying β_π and κ_π separately while keeping the graph, problem instance, initialization, and algorithm parameters fixed. Panel (a) illustrates how changing the communication weights alters either the mixing strength or the Perron distribution. Panels (b,c) show that weaker mixing and greater Perron imbalance both increase late-iteration KKT residuals when the other factor is fixed. These results support the separate roles of mixing and imbalance in Theorem 1. Experimental details are given in Appendix G.

Real-data learning. We evaluate the methods on two classification tasks: Adult (a9a), which predicts whether an individual's annual income exceeds \$50,000 from census attributes, and Multiple Features, which recognizes handwritten digits from six sets of extracted features. These correspond to binary and ten-class classification, respectively. Figures 4 and 5 show validation cross-entropy under communication budgets of $S = 2000$ and $S = 4000$, with results averaged over five paired seeds. Tables 2 and 3 summarize endpoint performance and gradient-query costs. On Adult, stochastic DiSPART achieves a validation cross-entropy of 0.3429, compared with 0.3287 for IPDC, while using approximately $1/40$ of the gradient queries consumed by either exact-gradient baseline. DiSPART-Det. achieves the smallest mean coupling-constraint violation, with an RMS residual of 4.72×10^{-5}. On Multiple Features, stochastic DiSPART attains a lower mean validation cross-entropy than all three baselines while using approximately $1/66$ of the queries required by either

exact-gradient baseline. DiSPART-Det. further achieves the lowest mean validation cross-entropy, coupling-constraint violation, and comparison residual among all methods. All runs satisfy the prescribed feasibility threshold $\mathrm{RMS} \leq 0.01$. Query costs include training and evaluation at the stated communication budgets.



(a) (b) (c)

Figure 3: **Effects of network mixing and Perron imbalance.** (a) Representative communication matrices W and Perron vectors π. (b) Mixing variation at fixed $\kappa_\pi = 10$. (c) Imbalance variation at fixed $\beta_\pi = 0.9$. Panel (a) uses five-node illustrations; panels (b,c) use ten-node networks with common algorithm parameters and paired oracle samples. Curves show means over ten seeds with pointwise 95% Student-t confidence bands.

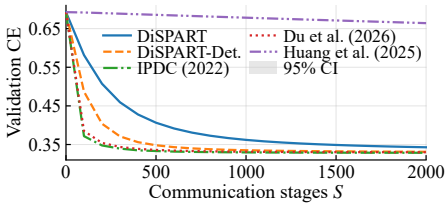


Figure 4: Val. CE on a9a.

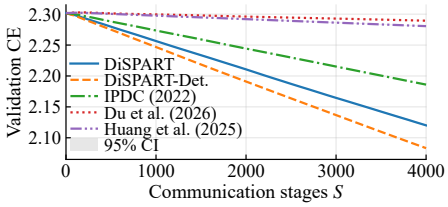


Figure 5: Val. CE on Multiple Features.

Table 2: a9a endpoints ($S = 2000$).

Method	Val. CE	RMS	R^{cmp}	$Q/10^6$
DiSPART	0.3429	5.97e−3	5.65e−4	1.315
DiSPART-Det.	0.3312	**4.72e−5**	4.23e−5	52.669
IPDC	**0.3287**	1.14e−3	**2.65e−6**	52.643
Du et al.	0.3299	3.55e−4	3.67e−5	52.643
Huang et al.	0.6642	7.68e−3	0.0791	**0.739**

Table 3: Multiple Features endpoints ($S = 4000$).

Method	Val. CE	RMS	R^{cmp}	$Q/10^6$
DiSPART	2.1201	9.13e−3	1.0015	0.073
DiSPART-Det.	**2.0832**	**2.48e−3**	**0.0745**	4.850
IPDC	2.1860	6.24e−3	0.4684	4.849
Du et al.	2.2894	2.52e−3	0.0769	4.849
Huang et al.	2.2806	4.98e−3	0.2988	**0.055**

Notes. CE: validation cross-entropy; RMS: root-mean-square coupling-constraint residual; R^{cmp}: comparison KKT residual; Q: cumulative sample-gradient-equivalent queries, including training and evaluation, at the stated communication budget (reported in millions). Lower values are better; bold marks column minima. Results are means over five paired seeds on a fixed split and network; shaded bands indicate pointwise 95% Student-t confidence intervals.

6 Conclusion

Distributed learning increasingly calls for methods that can handle stochastic nonconvex objectives, heterogeneous local variables, and asymmetric communication without relying on consensus alone. Motivated by this setting, we developed DiSPART for problems with general affine coupling over directed networks, combining stochastic gradient estimation, primal–dual stabilization, and distributed tracking of global constraint information. The analysis makes explicit how affine coupling, network mixing, and Perron imbalance enter the finite-time behavior, while the experiments support the proposed components, comparative performance, and the distinct effects of network topology. Extending these ideas to broader coupling models, time-varying or asynchronous networks, and larger-scale learning applications remains an important direction for future work.

AI USE STATEMENT

We used generative AI tools to assist with theoretical formulation and proof drafting and revision, experimental design, implementation and debugging of experimental and data-preparation code, literature search, reference checking, and manuscript editing and LaTeX preparation. Automated algebraic and software checks were used as supporting diagnostics during development. The authors are responsible for the validity of the mathematical arguments, experimental results, and references, and take responsibility for the final content of this work, including material produced with the assistance of generative AI.

ETHICS STATEMENT

This work focuses on methodological and theoretical advances in distributed optimization, supported by numerical experiments. The results should be interpreted under the assumptions and experimental settings stated in the paper. Practical applications should account for the ethical implications of the data and deployment context, including privacy, fairness, and security.

REPRODUCIBILITY STATEMENT

The problem formulation and assumptions are specified in Section 2, and Algorithm 1 gives the initialization and update rules. The supporting analysis and proofs of the main theorem and corollaries are provided in Appendices A–F. Appendix G documents the synthetic problem construction, algorithm and baseline configurations, network settings, evaluation criteria, resource accounting, and statistical reporting procedures, with implementation and randomization details in Appendix G.6.

REFERENCES

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 344–353. PMLR, 2019.

Kaixin Du, Min Meng, Xiuxian Li, and Keyou You. Nonconvex distributed composite optimization with coupled inequality constraints. *IEEE Transactions on Cybernetics*, 56(7):3677–3688, 2026. doi: 10.1109/TCYB.2026.3657045.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1529–1538. PMLR, 2017.

Ruichuan Huang, Jiawei Zhang, and Ahmet Alacaoglu. Stochastic smoothed primal-dual algorithms for nonconvex optimization with linear inequality constraints. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26106–26142. PMLR, 2025.

Liyuan Liang, Xinmeng Huang, Ran Xin, and Kun Yuan. Understanding the influence of digraphs on decentralized optimization: Effective metrics, lower bound, and optimal algorithm. *SIAM Journal on Optimization*, 35(3):1570–1600, 2025. doi: 10.1137/24M1657341.

Shivangi Dubey Sharma, Basil M. Idrees, Lavish Arora, and Ketan Rajawat. Decentralized stochastic constrained optimization via prox-linearization. arXiv preprint arXiv:2601.20345, 2026. URL `https://arxiv.org/abs/2601.20345`.

Haoran Sun, Songtao Lu, and Mingyi Hong. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9217–9228. PMLR, 2020.

Pedro Valdeira, Shiqiang Wang, and Yuejie Chi. Vertical federated learning with missing features during training and inference. In *International Conference on Learning Representations*, 2025.

Zesen Wang, Jiaojiao Zhang, Xuyang Wu, and Mikael Johansson. From promise to practice: Realizing high-performance decentralized training. In *International Conference on Learning Representations*, 2025.

Ran Xin, Usman Khan, and Soummya Kar. A hybrid variance-reduced method for decentralized stochastic non-convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11459–11469. PMLR, 2021.

Demyan Yarmoshik, Alexander Rogozin, Nikita Kiselev, Daniil Dorin, Alexander Gasnikov, and Dmitry Kovalev. Decentralized optimization with coupled constraints. In *International Conference on Learning Representations*, 2025.

Jiawei Zhang and Zhi-Quan Luo. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *SIAM Journal on Optimization*, 32(3):2319–2346, 2022. doi: 10.1137/20M135474X.

Jiawei Zhang, Songyang Ge, Tsung-Hui Chang, and Zhi-Quan Luo. Decentralized non-convex learning with linearly coupled constraints: Algorithm designs and application to vertical learning problem. *IEEE Transactions on Signal Processing*, 70:3312–3327, 2022. doi: 10.1109/TSP.2022.3184772.