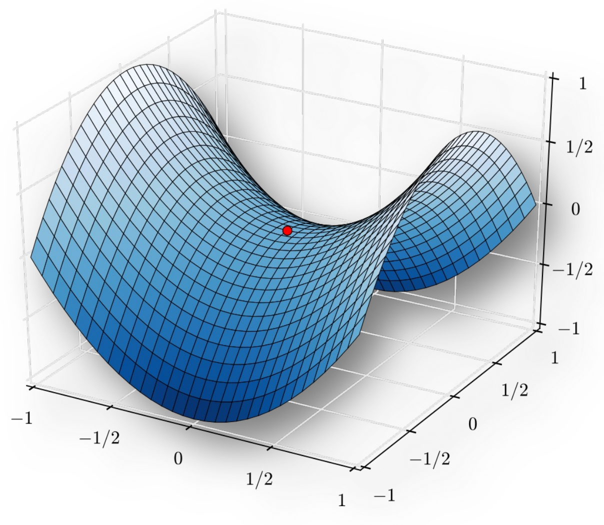
Local Composite Saddle Point Optimization


Site Bai
bai123@purdue.edu

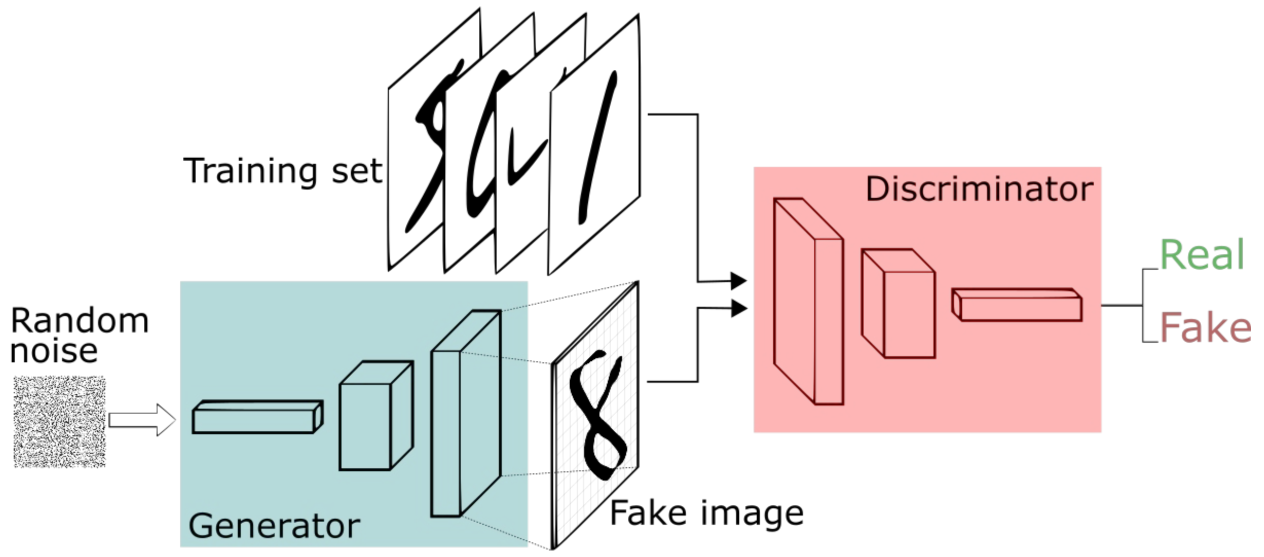
Brian Bullins
bbullins@purdue.edu

PURDUE UNIVERSITY
Department of Computer Science
West Lafayette, IN, USA

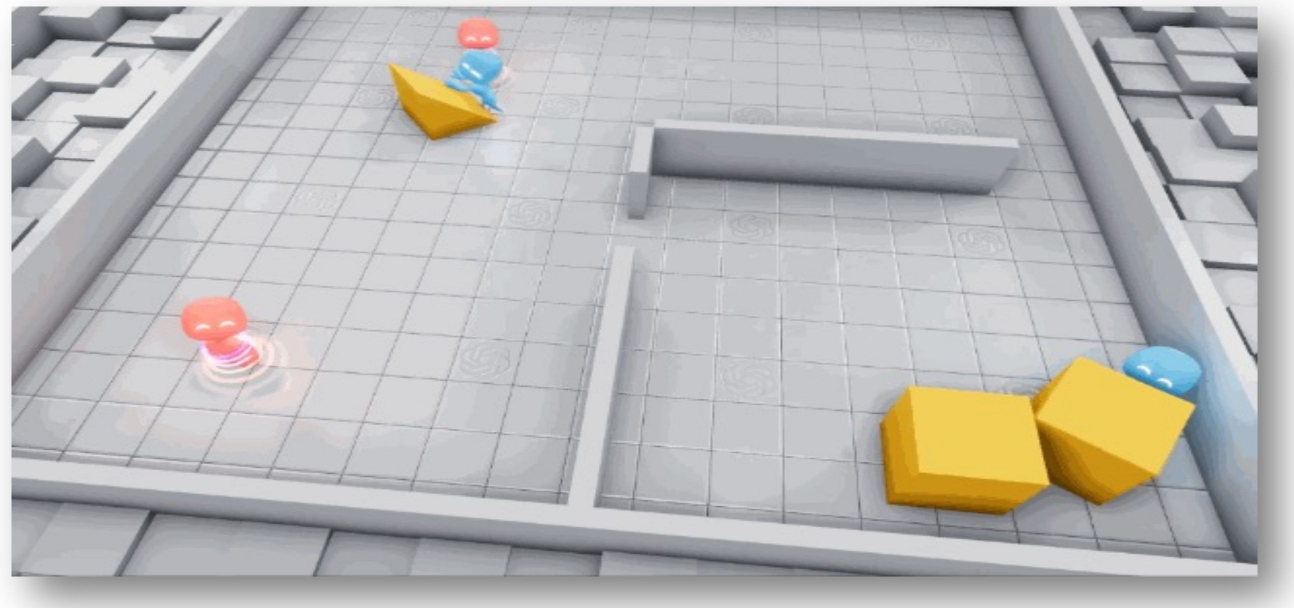
Saddle Point Optimization



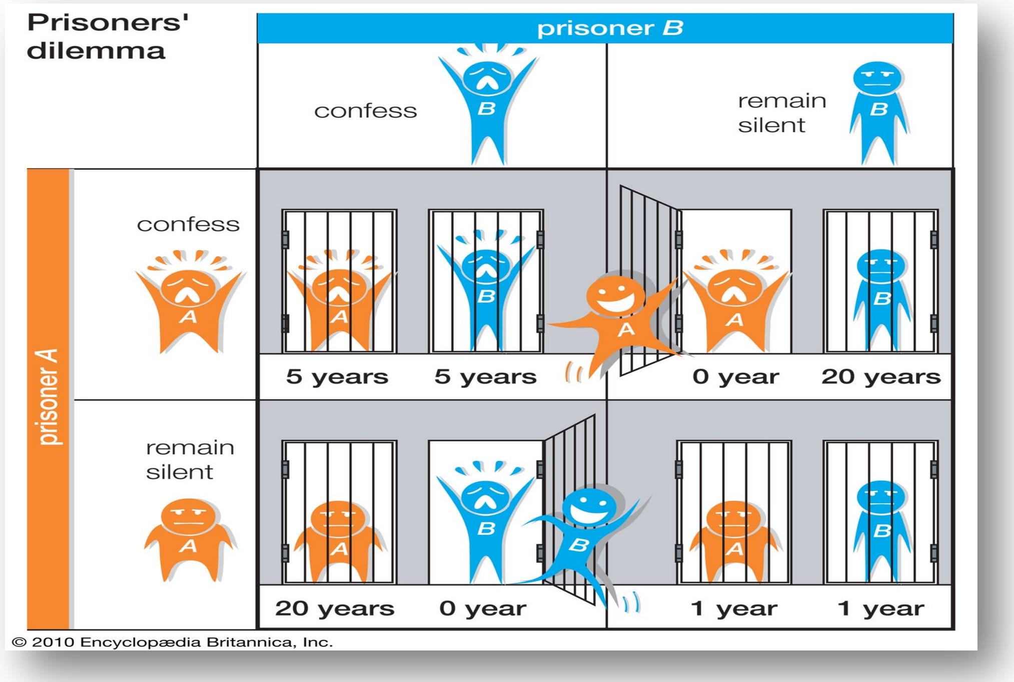
- Objective: $$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y)$$
- Applications:
  - Generative Adversarial Networks (GANs)



  - Multi-agent Reinforcement Learning



  - Matrix Games



  - More …

$$x_t = \mathrm{Prox}^h_{\bar{x}}(\mu_t)$$
$$x_{t+1/2} = \mathrm{Prox}^h_{x_t}(\eta g(x_t))$$
$$\mu_{t+1} = \mu_t + \eta g(x_{t+1/2})$$

Figure 1: Dual Extrapolation.

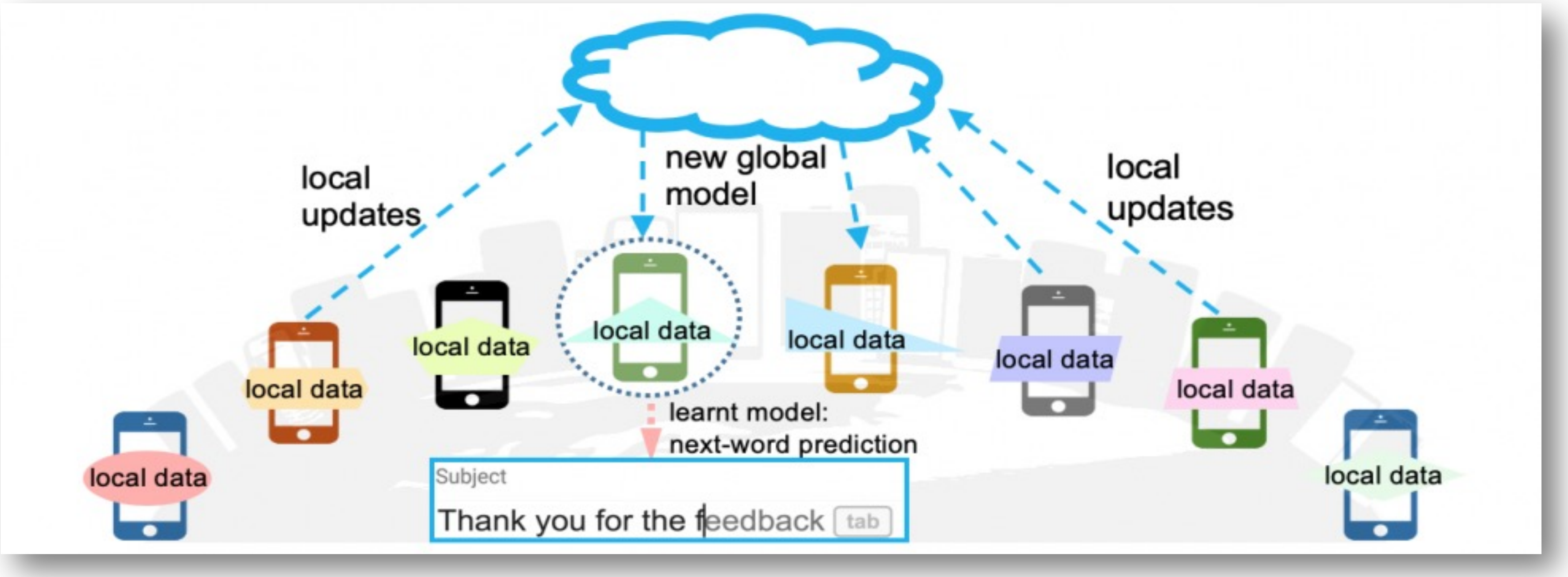
- Algorithms:
  - Nemirovski's Mirror Prox
  - Nesterov's Dual Extrapolation

proximal operator $\mathrm{Prox}^h_{x'}(\cdot) = \arg\min_x\{\langle \cdot, x\rangle + V^h_{x'}(x)\}$

Bregman divergence $V^h_{x'}(x) = h(x) - h(x') - \langle \nabla h(x'), x - x'\rangle$

Distributed Optimization / Federated Learning

- Federated Averaging / Local SGD
  - A server coordinates collaborative learning among clients
  - Cost of communication dominates the learning process
  - Local updates to improve communication efficiency
  - Aggregates local models through averaging



Algorithm 0 Typical FL Procedure

```
for r = 0, 1, ..., R − 1 do
    Sample a subset of clients
    Distribute global model to clients
    for each client in parallel do
        for k = 0, 1, ..., K − 1 do
            Certain optimization update
        end for
        Send local model to the server
    end parallel for
    Server aggregates client models
end for
```

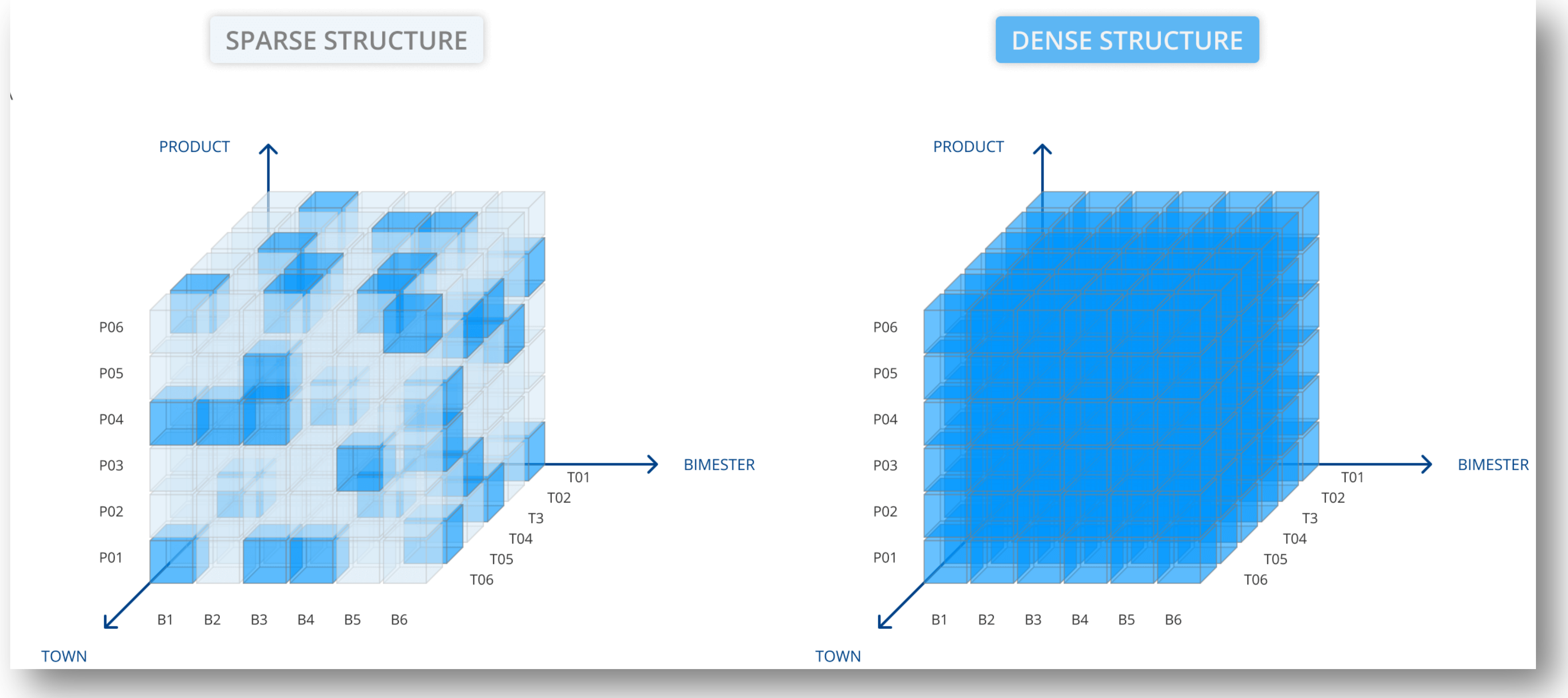
- Distributed Saddle Point Optimization [Beznosikov et al., 2020; Hou et al., 2021]
  - Objective: $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) = \frac{1}{M} \sum_{m=1}^{M} f_m(x, y)$

Composite Optimization / Non-smooth Regularization

- Objective: $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \phi(x,y) = f(x,y) + \psi_1(x) - \psi_2(y)$ where $f(x,y) = \frac{1}{M}\sum_{m=1}^{M} f_m(x,y)$
- Examples:

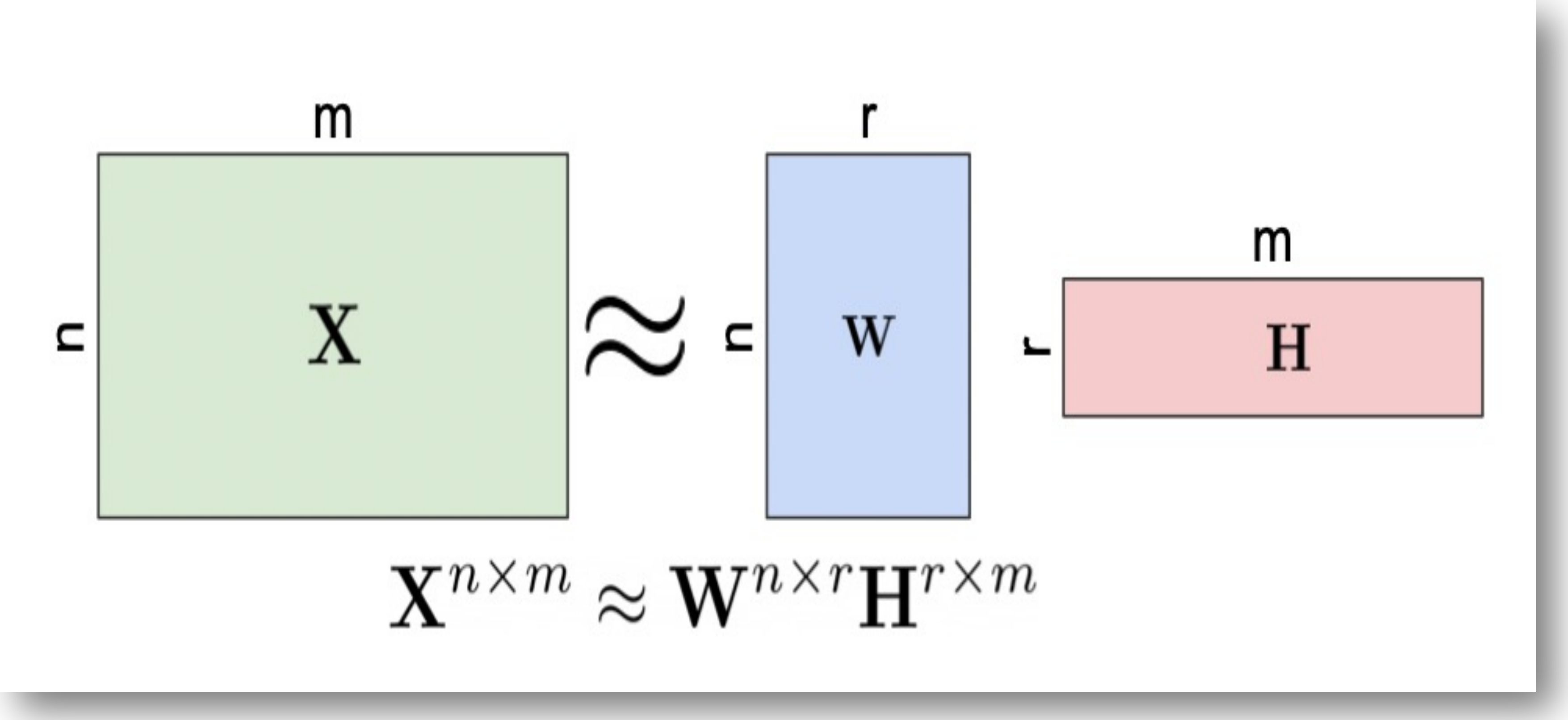
  - L_1 Regularization for sparsity

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle + \lambda\|\mathbf{x}\|_1 - \lambda\|\mathbf{y}\|_1$$



  - Nuclear norm regularization for low-rankness

$$\min_{\mathbf{X}\in\mathcal{X}}\max_{\mathbf{Y}\in\mathcal{Y}} \mathrm{Tr}\left((\mathbf{A}\mathbf{X} - \mathbf{B})^\top \mathbf{Y}\right) + \lambda\|\mathbf{X}\|_* - \lambda\|\mathbf{Y}\|_*$$



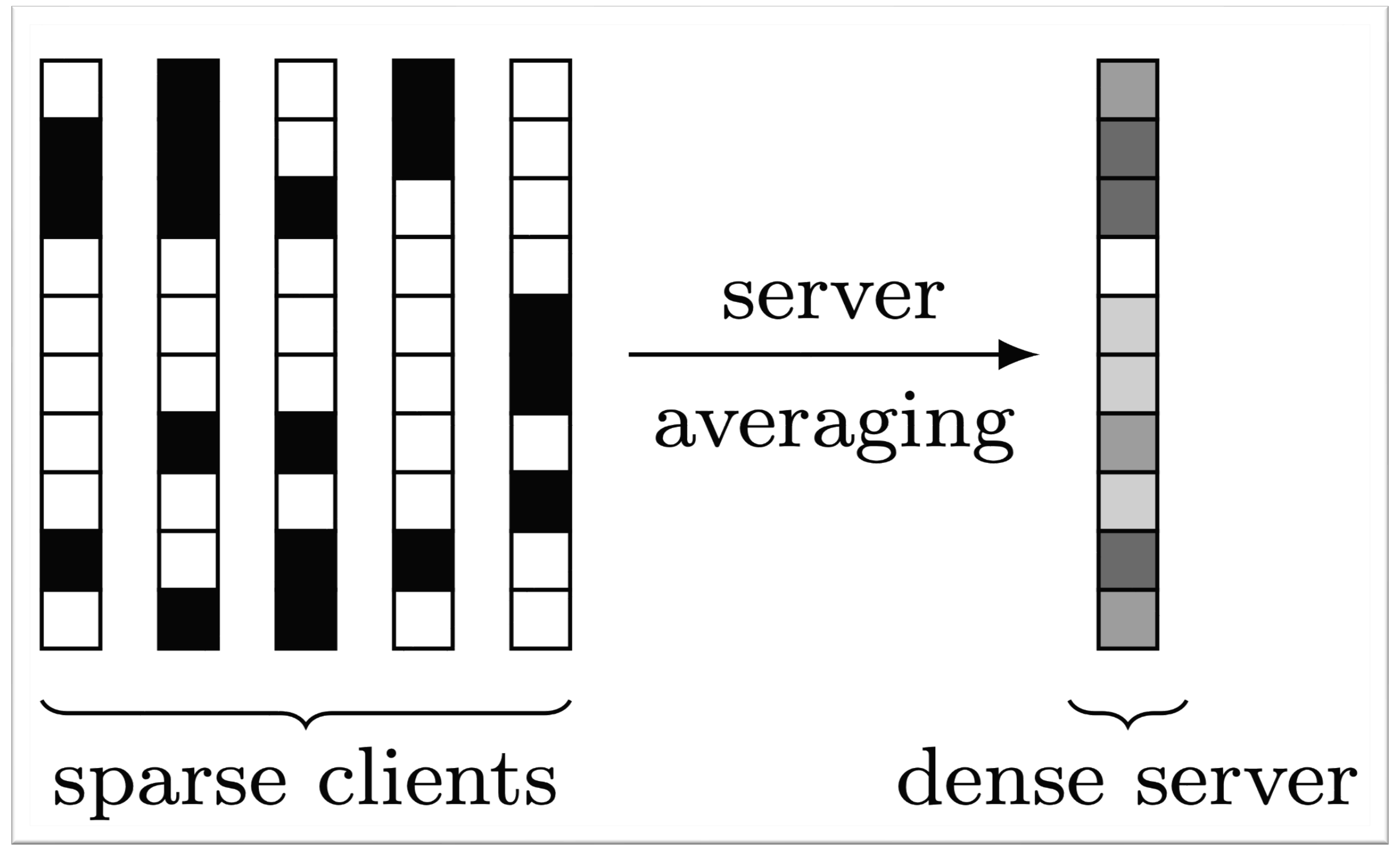
  - Indicator function for constraints

$$\psi(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases}$$

- None of existing distributed saddle point optimization algorithms can solve **composite objectives or objectives with constraints**

Federated Composite Optimization [Yuan et al., 2021]

- Curse of Primal Averaging in Federated Composite Optimization
  - Specific regularization-imposed structure on the clients no longer holds after direct averaging on the server
  - E.g. each client obtains a sparse solution, yet averaging the solutions across clients yields a dense solution
  - Propose Federated Dual Averaging that aggregates the dual solutions before projection to the primal space



$$\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (z^m_{r,K} - z^m_{r,0})$$ → Average client **dual** deltas

$$z_{r+1} \leftarrow z_r + \eta_s \Delta_r$$ → Server **dual** update

$$w_{r+1} \leftarrow \nabla(h + \eta_s \eta_c (r+1) K \psi)^*(z_{r+1})$$ → (Optional) primal output

- Federated Dual Averaging: Inferior Convergence for **Saddle Point Optimization**
  - Single-step Methods (Mirror Desenct / Dual Averaging [Bubeck et al., 2015]): $\mathcal{O}(1/\sqrt{T})$
  - Extra-step Methods (Mirror Prox [Nemirovski, 2004] / Dual Extrapolation[Nesterov, 2007]): $\mathcal{O}(1/T)$

Federated Dual Extrapolation (FeDualEx)

Task	Method	Composite & Constrained & Non-Euclidean
Min	FedAvg (Khaled et al., 2020)	✗
	FedDualAvg (Yuan et al., 2021)	✓
	FeDualEx (Ours)	✓
Min-Max	Extra Step Local SGD (Beznosikov et al., 2020)	✗
	SCCAFFOLD-S (Hou et al., 2021)	✗
	FeDualEx (Ours)	✓

- Present the first algorithm for saddle point optimization with composite non-smooth regularization under a distributed paradigm, and derive its convergence rate
- Showcase the structure-preserving (e.g., sparsity) advantage of FeDualEx achieved through dual-space averaging
- Present deterministic and stochastic dual extrapolation for composite saddle point optimization in the sequential setting
- Demonstrate experimentally the effectiveness of FeDualEx on various composite saddle point tasks

Federated Dual Extrapolation (FeDualEx)

Definition 3 (Generalized Bregman Divergence for Saddle Functions). *The generalized distance-generating function for the optimization of* (1) *is* $\ell_t(z) = \ell(z) + t\eta\psi(z)$, *where* $\ell(z) = h_1(x) + h_2(y)$, h_1 *and* h_2 *are distance-generating functions for* x *and* y, $\psi(z) = \psi_1(x) + \psi_2(y)$, η *is the step size, and* t *is the current number of iterations. It generates the following generalized Bregman divergence:*

$$\tilde{V}^{\ell_t}_{\varsigma'}(z) = \ell_t(z) - \ell_t(z') - \langle \varsigma', z - z' \rangle,$$

where ς' *is the preimage of* z' *with respect to the gradient of the conjugate of* ℓ_t, *i.e.,* $z' = \nabla \ell_t^*(\varsigma')$.

Definition 4 (Generalized Proximal Operator for Saddle Functions). *A proximal operation in the composite setting with generalized Bregman divergence for Saddle Functions is defined to be*

$$\tilde{\mathrm{Prox}}^{\ell_t}_{\varsigma'}(g) := \arg\min_z \{\langle g, z \rangle + \tilde{V}^{\ell_t}_{\varsigma'}(z)\},$$

where ς' *is the dual image of* z', *i.e.,* $z' = \nabla \ell_t^*(\varsigma')$, *and* $\varsigma' \in \partial \ell_t(z') = \nabla \ell(z') + \eta t \partial \psi(z')$.

Federated Dual Extrapolation (FeDualEx)

Algorithm 1 FEDERATED-DUAL-EXTRAPOLATION (FeDualEx) for Composite SPP

Input: $\phi(z) = f(x,y) + \psi_1(x) - \psi_2(y) = \frac{1}{M}\sum_{m=1}^{M} f_m(x,y) + \psi_1(x) - \psi_2(y)$: objective function; $\ell(z)$: distance-generating function; $g_m(z) = (\nabla_x f_m(x,y), -\nabla_y f_m(x,y))$: gradient operator.

Hyperparameters: R: number of communication rounds; K: number of local update iterations; η^s: server step size; η^c: client step size.

Dual Initialization: $\varsigma_0 = 0$: initial dual variable, $\bar{\varsigma}$: fixed point in the dual space.

Output: Approximate solution $z = (x, y)$ to $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \phi(x,y)$

for $r = 0, 1, \ldots, R-1$ **do**
    Sample a subset of clients $C_r \subseteq [M]$
    for $m \in C_r$ **in parallel do**
        $\varsigma_{r,0}^m = \varsigma_r$
        for $k = 0, 1, \ldots, K-1$ **do**
            $z_{r,k}^m = \widetilde{\mathrm{Prox}}_{\bar{\varsigma}}^{\ell_{r,k}}(\varsigma_{r,k}^m)$ ▷ Two-step evaluation of the generalized proximal operator
            $z_{r,k+1/2}^m = \widetilde{\mathrm{Prox}}_{\bar{\varsigma}-\varsigma_{r,k}^m}^{\ell_{r,k+1}}(\eta^c g_m(z_{r,k}^m; \xi_{r,k}^m))$
            $\varsigma_{r,k+1}^m = \varsigma_{r,k}^m + \eta^c g_m(z_{r,k+1/2}^m; \xi_{r,k+1/2}^m)$ ▷ Dual variable update
        end for
    end parallel for
    $\Delta_r = \frac{1}{|\mathcal{C}_r|}\sum_{m\in\mathcal{C}_r}(\varsigma_{r,K}^m - \varsigma_{r,0}^m)$
    $\varsigma_{r+1} = \varsigma_r + \eta^s \Delta_r$ ▷ Server dual update
end for
Return: $\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1} \widehat{z_{r,k+1/2}}$ with $\widehat{z_{r,k+1/2}}$ defined in (4).

Assumptions *For the composite saddle function* $\phi(x,y) = \frac{1}{M}\sum_{m=1}^{M} f_m(x,y) + \psi_1(x) - \psi_2(y)$, *its gradient operator is given by* $g = (\nabla_x f, -\nabla_y f)$ *and* $g = \frac{1}{M}\sum_{m=1}^{M} g_m$. *We assume that*

a. *(Convexity of* f*)* $\forall m \in [M]$, $f_m(x,y)$ *is convex in* x *and concave in* y.

b. *(Convexity of* ψ*)* $\psi_1(x)$ *is convex in* x, *and* $\psi_2(y)$ *is convex in* y.

c. *(Lipschitzness of* g*)* $g_m(z) = \begin{bmatrix} \nabla_x f_m(x,y) \\ -\nabla_y f_m(x,y) \end{bmatrix}$ *is* β*-Lipschitz:* $\|g_m(z) - g_m(z')\|_* \le \beta\|z - z'\|$

d. *(Unbiased Estimate and Bounded Variance)* $\forall m \in [M]$, *for random sample* ξ^m, $\mathbb{E}_\xi[g_m(z^m;\xi^m)] = g_m(z^m)$, *and* $\mathbb{E}_\xi\left[\|g_m(z^m;\xi^m) - g_m(z^m)\|_*^2\right] \le \sigma^2$

e. *(Bounded Gradient)* $\forall m \in [M]$, $\|g_m(z^m;\xi^m)\|_* \le G$

f. *The distance-generating function* ℓ *is a Legendre function that is 1-strongly convex, i.e.,* $\forall z, z'$,

$$\ell(z') - \ell(z) - \langle \nabla \ell(z), z' - z\rangle \ge \frac{1}{2}\|z' - z\|^2.$$

g. *The optimization domain* $\mathcal{Z}$ *is compact w.r.t. Bregman divergence, i.e.,* $\forall z, z' \in \mathcal{Z}$, $V_{z'}^{\ell}(z) \le B$.

Theorem 1 (Main). *Under assumptions, the duality gap evaluated with the ergodic sequence generated by the intermediate steps of FeDualEx in Algorithm 1 is bounded by*

$$\mathbb{E}\left[\operatorname{Gap}\left(\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1} \widehat{z_{r,k+1/2}}\right)\right] \le \frac{5^{\frac{1}{2}}\beta B}{RK} + \frac{20^{\frac{1}{4}}\beta^{\frac{1}{2}}G^{\frac{1}{2}}B^{\frac{3}{4}}}{K^{\frac{1}{4}}R^{\frac{3}{4}}} + \frac{5^{\frac{1}{2}}\sigma B^{\frac{1}{2}}}{M^{\frac{1}{2}}R^{\frac{1}{2}}K^{\frac{1}{2}}} + \frac{2^{\frac{3}{4}}\beta^{\frac{1}{2}}G^{\frac{1}{2}}B^{\frac{3}{4}}}{R^{\frac{1}{2}}}.$$

Other Settings

Distributed Composite Convex Optimization [Yuan et al., 2021]

Theorem 2. *Under the convex counterparts of previous assumptions, choosing step size* $\eta^c = \min\{\frac{1}{5^{\frac{1}{2}}\beta}, \frac{B^{\frac{1}{4}}}{20^{\frac{1}{4}}\beta^{\frac{1}{2}}G^{\frac{1}{2}}K^{\frac{3}{4}}R^{\frac{1}{4}}}, \frac{B^{\frac{1}{2}}M^{\frac{1}{2}}}{5^{\frac{1}{2}}\sigma R^{\frac{1}{2}}K^{\frac{1}{2}}}, \frac{B^{\frac{1}{3}}}{2^{\frac{1}{3}}\beta^{\frac{1}{3}}G^{\frac{2}{3}}KR^{\frac{1}{3}}}\}$, *the ergodic intermediate sequence generated by FeDualEx for composite convex objectives satisfies*

$$\mathbb{E}[\phi(\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\widehat{x_{r,k+1/2}}) - \phi(x)] \leq \frac{5^{\frac{1}{2}}\beta B}{RK} + \frac{20^{\frac{1}{4}}\beta^{\frac{1}{2}}G^{\frac{1}{2}}B^{\frac{3}{4}}}{K^{\frac{1}{4}}R^{\frac{3}{4}}} + \frac{5^{\frac{1}{2}}\sigma B^{\frac{1}{2}}}{M^{\frac{1}{2}}R^{\frac{1}{2}}K^{\frac{1}{2}}} + \frac{2^{\frac{1}{3}}\beta^{\frac{1}{3}}G^{\frac{2}{3}}B^{\frac{2}{3}}}{R^{\frac{2}{3}}}.$$

Stochastic Composite Saddle Point Optimization [Mishchenko et al., 2020]

Theorem 3. *Under the sequential versions of previous assumptions,* $\forall z \in \mathcal{Z}$, *choosing step size* $\eta = \min\{\frac{1}{3^{\frac{1}{2}}\beta}, \frac{B^{\frac{1}{2}}}{3^{\frac{1}{2}}\sigma T^{\frac{1}{2}}}\}$, *the ergodic intermediate sequence of stochastic dual extrapolation satisfies* $\mathbb{E}\big[\operatorname{Gap}(\frac{1}{T}\sum_{t=0}^{T-1} z_{t+1/2})\big] \leq \frac{3^{\frac{1}{2}}\beta B}{T} + \frac{3^{\frac{1}{2}}\sigma B^{\frac{1}{2}}}{T^{\frac{1}{2}}}$.

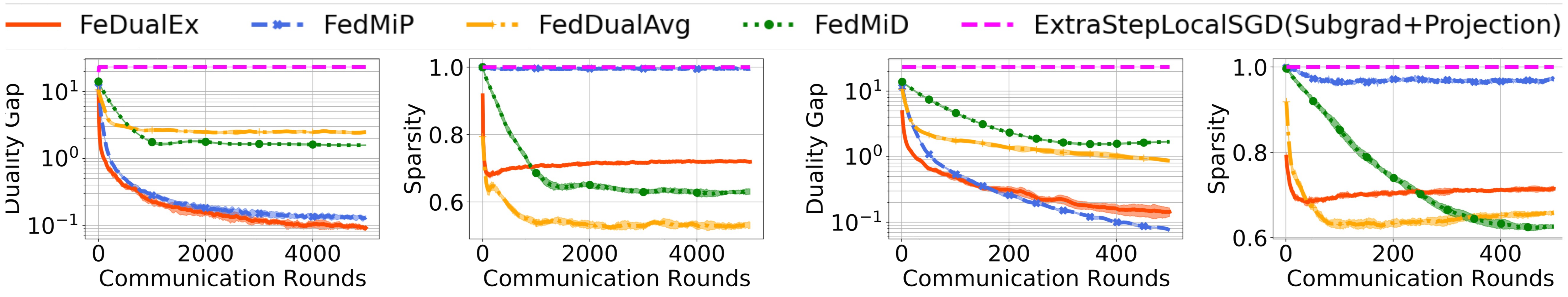
Deterministic Composite Saddle Point Optimization [He et al., 2015]

Theorem 4. *Under the basic convexity assumption and* β*-Lipschitzness of* g, $\forall z \in \mathcal{Z}$ *and* $\eta \leq \frac{1}{\beta}$, *composite dual extrapolation satisfies* $\operatorname{Gap}(\frac{1}{T}\sum_{t=0}^{T-1} z_{t+1/2}) \leq \frac{\beta B}{T}$.

Composite Bilinear Saddle Point Problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle + \lambda\|\mathbf{x}\|_1 - \lambda\|\mathbf{y}\|_1$$

$$\mathbf{A} \in \mathbb{R}^{n\times m}, \quad \mathcal{X} = \{\mathbb{R}^m : \|\mathbf{x}\|_\infty \le D\},$$
$$\mathbf{b} \in \mathbb{R}^n, \quad \mathcal{Y} = \{\mathbb{R}^n : \|\mathbf{y}\|_\infty \le D\}.$$



(a) One Local Update ($K = 1$)

(b) Ten Local Updates ($K = 10$)

Figure 4: Duality gap and sparsity of the solution for ℓ_1 regularized SPP with ℓ_∞ constraint.

$$\min_{\mathbf{X}\in\mathcal{X}} \max_{\mathbf{Y}\in\mathcal{Y}} \mathrm{Tr}\big((\mathbf{A}\mathbf{X} - \mathbf{B})^\top \mathbf{Y}\big) + \lambda\|\mathbf{X}\|_* - \lambda\|\mathbf{Y}\|_*$$

$$\mathbf{A} \in \mathbb{R}^{n\times m}, \quad \mathcal{X} = \{\mathbb{R}^{m\times p} : \|\mathbf{X}\|_2 \le D\},$$
$$\mathbf{B} \in \mathbb{R}^{n\times p}, \quad \mathcal{Y} = \{\mathbb{R}^{n\times p} : \|\mathbf{Y}\|_2 \le D\}.$$

(a) $K = 1$

(b) $K = 10$

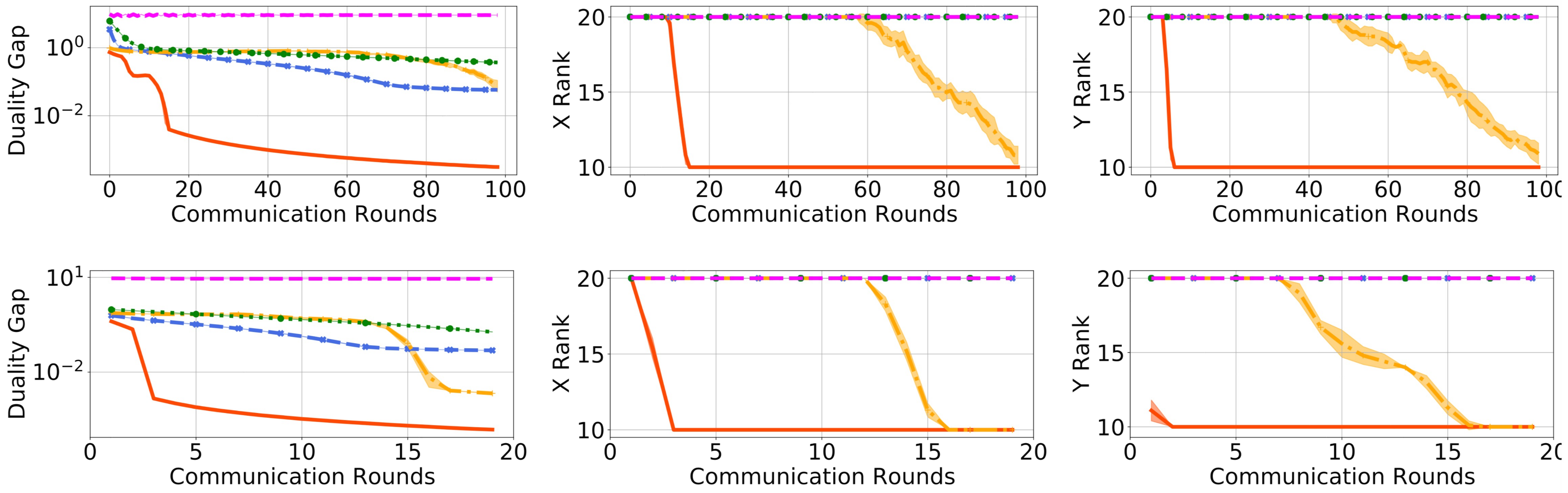


Figure 5: Duality gap and rank of the solution to the nuclear norm regularized SPP.

Universal Adversarial Training of Logistic Regression

$$\min_{\mathbf{w}\in\mathbb{R}^d} \max_{\|\boldsymbol{\delta}\|_\infty \le D} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^\top(\mathbf{x}_i+\boldsymbol{\delta}), y_i) + \lambda\|\boldsymbol{\delta}\|_1$$

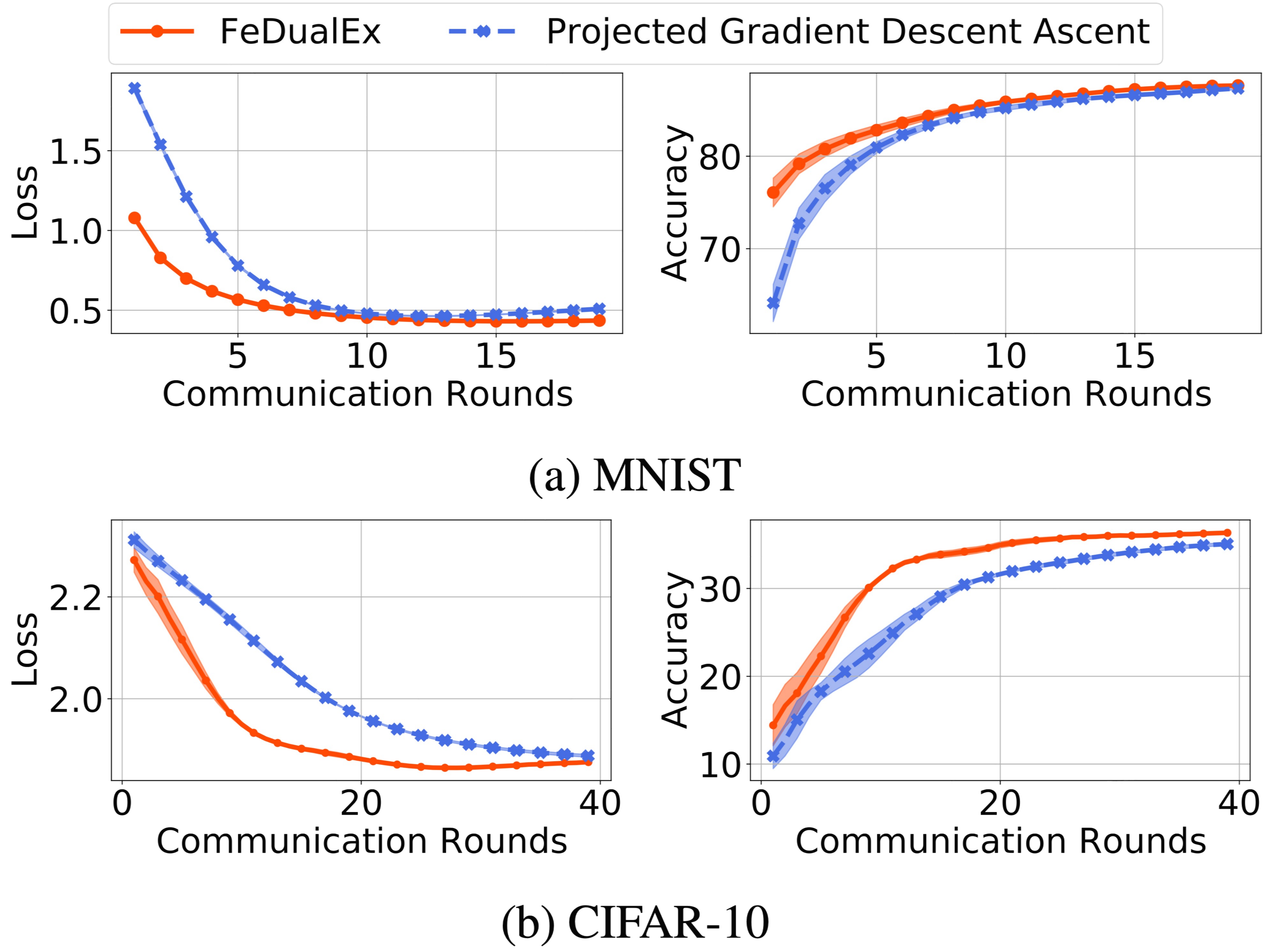


Figure 6: Universal adversarial training loss and validation accuracy of logistic regression on unattacked data.

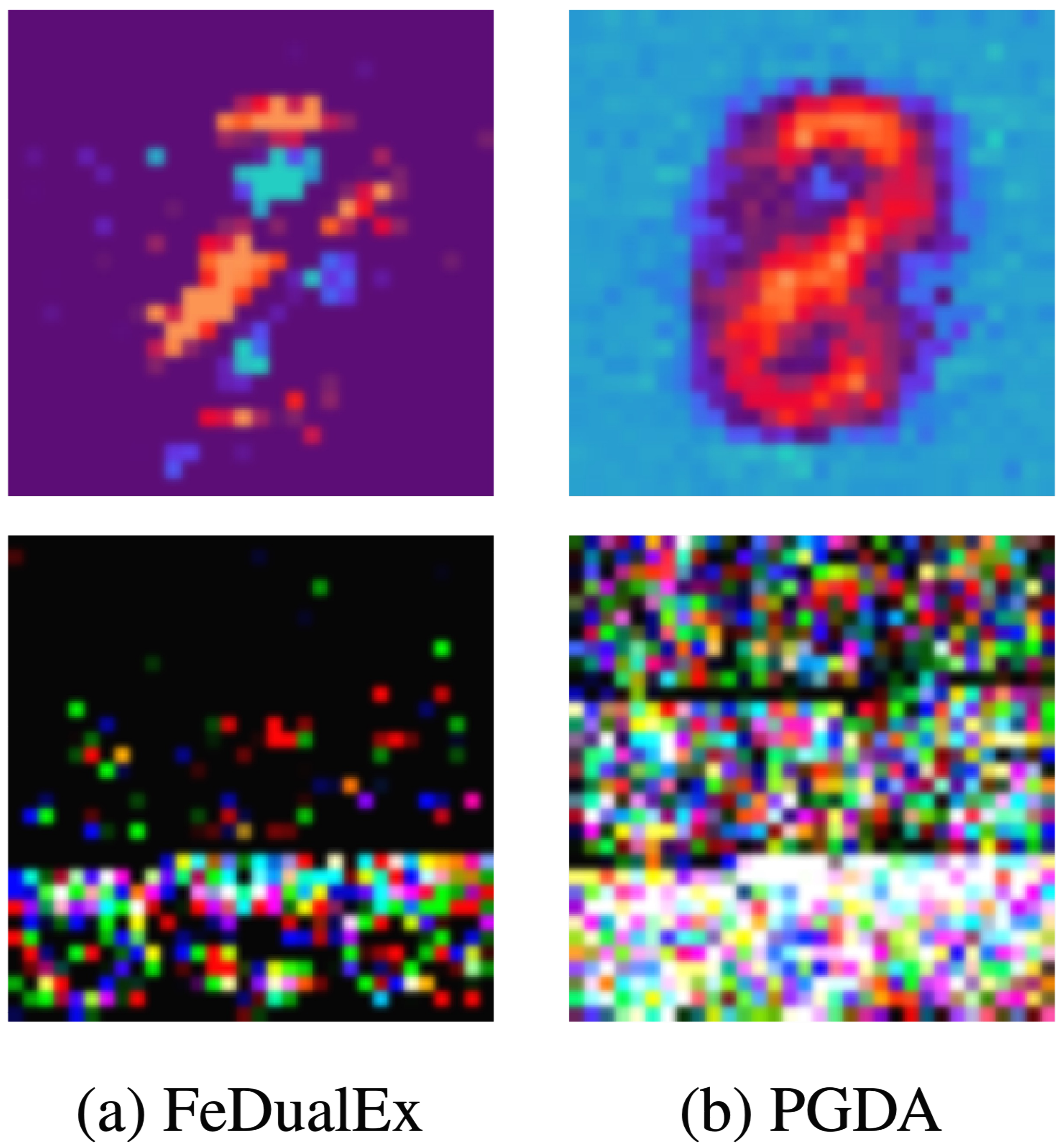


Figure 7: Attack generated from the universal-adversarially trained logistic regression on MNIST and CIFAR-10.

Local Composite Saddle Point Optimization

References:

Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle-point problems: Lower bounds, optimal and robust algorithms. arXiv preprint arXiv:2010.13112, 2020.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. Foundations and Trends® in Machine Learning, 8(3-4):231–357, 2015.

Niao He, Anatoli Juditsky, and Arkadi Nemirovski. Mirror prox algorithm for multi-term composite minimization and semi-separable problems. Computational Optimization and Applications, 61: 275–319, 2015.

Charlie Hou, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Efficient algorithms for federated saddle point optimization. arXiv preprint arXiv:2102.06333, 2021.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In International Conference on Artificial Intelligence and Statistics, pp. 4573–4582. PMLR, 2020.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM Journal on Optimization, 15(1):229–251, 2004.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. Mathematical Programming, 109(2-3):319–344, 2007.

Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. In International Conference on Machine Learning, pp. 12253–12266. PMLR, 2021.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In International Conference on Artificial Intelligence and Statistics, pp. 4519–4529. PMLR, 2020.