Distributed Saddle-Point Problems: Lower Bounds, Optimal Algorithms and Federated GANs

Aleksandr Beznosikov [1,2,4], Valentin Samokhin [1,3], Alexander Gasnikov [1,2,3,4]

[1] MIPT(Russia), [2] HSE(Russia), [3] IITP(Russia), [4] Sirius(Russia)

1. The Problem

Problem. Distributed saddle-point problem:

$$\min_{x\in X}\max_{y\in Y} f(x,y) := \frac{1}{M}\sum_{m=1}^{M} f_m(x,y).$$

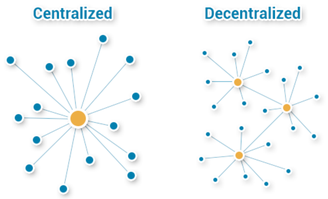
Assumptions. Stochastic gradients, Lipschitz continuity, (strong-)convexity-(strong-)concavity

$$F_m(z) = F_m(x,y) = \begin{pmatrix} \nabla_x f_m(x,y) \\ -\nabla_y f_m(x,y) \end{pmatrix} \qquad \text{only stochastic } F_m(z,\xi)$$

$$\|F(z_1)-F(z_2)\| \le L\|z_1-z_2\| \qquad \langle F(z_1)-F(z_2), z_1-z_2\rangle \ge \mu\|z_1-z_2\|^2$$

$$\mathbb{E}[F_m(z,\xi)] = F_m(z),\ \mathbb{E}[\|F_m(z,\xi)-F_m(z)\|^2] \le \sigma^2 \qquad \|z-z'\| \le \Omega_z$$

Centalized and Decenralized cases.



communication graph $G(\mathcal{V},\mathcal{E})$ with diameter Δ and char. number

Comminication is a bottleneck!

✓ In the case of **convex optimization**, lower bounds for centralized and decentralized alogorithms were proved [5], optimal methods were obtained [6], and the Local SGD technique was developed in a large number of works [2,3].

✗ **Saddle point problems** are a more complex and no less interesting class of problems. At the same time, the development of the theory of distributed learning is almost absent.

✗ **Local SGD** techniques proposed in the literature in recent months [1] work with the number of iterations 10^9.

2. Our contributions

✓ **Lower bounds for distributed saddle-poinr problems** in centralized and decentralized cases with fixed number of communications and local iterations

✓ **Optimal algorithms** in centralized and decentralized cases, which reaches the lower bounds

✓ **Local SGD type method for saddle-point problems.** The method step depends on the Lipschitz constant and the communication frequency in the first degree.

✓ **Federated GANs** training on MNIST with Local SGD type methods. Experiments are carried out on highly heterogeneous data with a small number of communications

3. Lower and upper bounds

Convergence in expectation of a distance to the solution:

$$\mathbb{E}[\|z-z^*\|^2] \qquad \text{or} \qquad \mathbb{E}[gap(z)]$$

$$gap(z) = \max_{y'\in\mathcal{Y}} f(x,y') - \min_{x'\in\mathcal{X}} f(x',y)$$

Here T - number of local calls in each device, K - number of communications. $R_0 = \|z_0 - z^*\|$

	lower	upper
	centralized	
sc	$\Omega\left(R_0^2\exp\left(-\frac{32\mu K}{L\Delta}\right)+\frac{\sigma^2}{\mu^2 MT}\right)$	$\tilde{\mathcal{O}}\left(R_0^2\exp\left(-\frac{\mu K}{4L\Delta}\right)+\frac{\sigma^2}{\mu^2 MT}\right)$
c	$\Omega\left(\frac{L\Omega_z^2\Delta}{K}+\frac{\sigma\Omega_z}{\sqrt{MT}}\right)$	$\mathcal{O}\left(\frac{L\Omega_z^2\Delta}{K}+\frac{\sigma\Omega_z}{\sqrt{MT}}\right)$
	decentralized	
sc	$\Omega\left(R_0^2\exp\left(-\frac{128\mu K}{L\sqrt{\chi}}\right)+\frac{\sigma^2}{\mu^2 MT}\right)$	$\tilde{\mathcal{O}}\left(R_0^2\exp\left(-\frac{\mu K}{8L\sqrt{\chi}}\right)+\frac{\sigma^2}{\mu^2 MT}\right)$
c	$\Omega\left(\frac{L\Omega_z^2\sqrt{\chi}}{K}+\frac{\sigma\Omega_z}{\sqrt{MT}}\right)$	$\tilde{\mathcal{O}}\left(\frac{L\Omega_z^2\sqrt{\chi}}{K}+\frac{\sigma\Omega_z}{\sqrt{MT}}\right)$

Bad functions for lower bounds - bilinear problem with block structure.

Optimal algorithms for upper bounds 1) in cenralized case - Extra Step Method with right batch size, 2) in decentralized case - combining of Extra Step Method and Accelerated Gossip(FastMix) [4].

Algorithm 1 Centralized Extra Step Method

Parameters: Stepsize $\gamma \le \frac{1}{4L}$;
Communication rounds K, number of local steps T.
Initialization: Choose $(x^0,y^0)=z^0\in\mathcal{Z}$,
$k=\lfloor\frac{K}{r}\rfloor$ and batch size $b=\lfloor\frac{T}{2k}\rfloor$.
for $t=0,1,2,\dots,k$ **do**
  for each machine m **do**
    $g_m^t = \frac{1}{b}\sum_{i=1}^{b} F_m(z^t,\xi_m^{t,i})$, send g_m^t,
  on server:
    $z^{t+1/2} = \text{proj}_{\mathcal{Z}}(z^t - \frac{\gamma}{M}\sum_{m=1}^{M} g_m^t)$, send $z^{t+1/2}$,
  for each machine m **do**
    $g_m^{t+1/2} = \frac{1}{b}\sum_{i=1}^{b} F_m(z^{t+1/2},\xi_m^{t+1/2,i})$, send $g_m^{t+1/2}$,
  on server:
    $z^{t+1} = \text{proj}_{\mathcal{Z}}(z^t - \frac{\gamma}{M}\sum_{m=1}^{M} g_m^{t+1/2})$, send z^{t+1},
end for
Output: z^{k+1} or z_{avg}^{k+1}.

Both algorithms use the operator of Euclidean projection: $\text{proj}_{\mathcal{Z}}(z) = \min_{u\in\mathcal{Z}}\|u-z\|$

Local devices send information to the server, it sends them a response.

Algorithm 2 Decentralized Extra Step Method

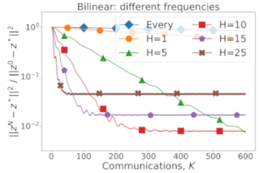
Parameters: Stepsize $\gamma \le \frac{1}{4L}$;
Communication rounds K, number of local calls T.
Initialization: Choose $(x^0,y^0)=z^0\in\mathcal{Z}$, $z_m^0=z^0$,
$k=\lfloor\frac{K}{H}\rfloor$ and batch size $b=\lfloor\frac{T}{2k}\rfloor$.
for $t=0,1,2,\dots,k$ **do**
  for each machine m **do**
    $g_m^t = \frac{1}{b}\sum_{i=1}^{b} F_m(z_m^t,\xi_m^{t,i})$, $\hat{z}_m^{t+1/2} = z_m^t - \gamma g_m^t$,
  communication
    $\hat{z}_1^{t+1/2},\dots,\hat{z}_M^{t+1/2} = \text{FastMix}(\hat{z}_1^{t+1/2},\dots,\hat{z}_M^{t+1/2}, H)$,
  for each machine m **do**
    $z_m^{t+1/2} = \text{proj}_{\mathcal{Z}}(\hat{z}_m^{t+1/2})$,
    $g_m^{t+1/2} = \frac{1}{b}\sum_{i=1}^{b} F_m(z_m^{t+1/2},\xi_m^{t+1/2,i})$,
    $\hat{z}_m^{t+1} = z_m^t - \gamma g_m^{t+1/2}$,
  communication
    $\hat{z}_1^{t+1},\dots,\hat{z}_M^{t+1} = \text{FastMix}(\hat{z}_1^{t+1},\dots,\hat{z}_M^{t+1}, H)$,
  for each machine m **do**
    $z_m^{t+1} = \text{proj}_{\mathcal{Z}}(\hat{z}_m^{t+1})$,
end for
Output: $\bar{z}^{k+1}$ or $\bar{z}_{avg}^{k+1}$.

H - number of iterations of Accelerated Gossip (FastMix) procedure. The output of it is approximately the same vectors.

4. Extra Step Local SGD

This method is similar to Algorithm 1, but it **does not communicate every iteration**. The convergence estimates for Algorithm 3 are not optimal, but in practice **it is better in terms of the number of communications**:

$$\min_{x,y\in[-1;1]^n}\max \frac{1}{M}\sum_{m=1}^{M}(x^T A_m y + b_m^T x + c_m^T y)$$



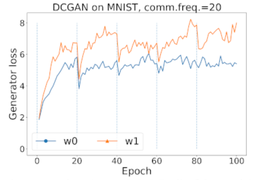
Algorithm 3 Extra Step Local SGD

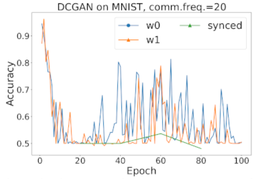
Parameters: stepsize $\gamma \le \frac{1}{6HL_{\max}}$;
number of local steps T,
sets I of communications steps for x and y ($|I|=K$).
Initialization: Choose $(x^0,y^0)=z^0\in\mathcal{Z}$,
for all m $z_m^0 = z^0$ and $\bar{z}=z^0$.
for $k=0,1,2,\dots,T$ **do**
  for each machine m **do**
    $z_m^{k+1/2} = \text{proj}_{\mathcal{Z}}(z_m^k - \gamma F_m(z_m^k,\xi_m^k))$,
    $z_m^{k+1} = \text{proj}_{\mathcal{Z}}(z_m^k - \gamma F_m(z_m^{k+1/2},\xi_m^{k+1/2}))$,
    if $k\in I$, send z_m^{k+1} on server,
  on server:
    if $k\in I$ compute $\bar{z} = \frac{1}{M}\sum_{m=1}^{M} z_m^{k+1}$, send $\bar{z}$.
  for each machine m **do**
    if $k\in I$, get $\bar{z}$ and set $z_m^{k+1} = \bar{z}$,
end for
Output: $\bar{z}$.

Here H - number of local steps without communications (frequency of communications).

5. Federated GANs

GAN - saddle-point problem. We trainn DCGAN on MNIST with Local SGD and Local ADAM technique. The pictures show the generator's loss and the discriminator's accuracy on two devices, depending on the epoch number (communications every 20 epochs):





An digit pictures by Local SGD (left) and Local Adam (right) technique on heterogeneous data (each of nodes has own set of unique digits). Over 100 epochs, we communicate only 5 times:

References

[1] Deng, Y., Mahdavi, M.: Local stochastic gradient descent ascent: Convergence analysis and communication effciency. In: Proceedings of The 24th International Conference on Artifcial Intelligence and Statistics. pp. 1387-1395 (2021)

[2] Khaled, A., Mishchenko, K., Richtarik, P.: Tighter theory for local SGD on identical and heterogeneous data. In: International Conference on Artifcial Intelligence and Statistics. pp. 4519-4529 (2020)

[3] Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., Stich, S.: A unifed theory of decentralized SGD with changing topology and local updates. In: Proceedings of the 37th International Conference on Machine Learning. pp. 5381-5393 (2020)

[4] Liu, J., Morse, A.S.: Accelerated linear iterations for distributed averaging. Annual Reviews in Control 35(2), 160-165 (2011)

[5] Scaman, K., Bach, F., Bubeck, S., Lee, YT., Massoulie, L.: Optimal algorithms for smooth and strongly convex distributed optimization in networks. In: international conference on machine learning. pp. 3027-3036 (2017)

[6] Ye, H., Zhou, Z., Luo, L., Zhang, T.: Decentralized accelerated proximal gradient descent. In: Advances in Processing Systems. vol. 33, pp. 18308-18317 (2020)