Stochastic zero order methods for unconstrained minimization

El Houcine Bergou (elhoucine.bergou@um6p.ma)

MOHAMMED VI POLYTECHNIC UNIVERSITY

UM6P-CS SCHOOL OF COMPUTER SCIENCE

November 2, 2022

Outline

The Problem

Consider the optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{1}$$

d: the number of parameters
n: the number of samples or the number of clients participating in the learning task
$f_i : \mathbb{R}^d \to \mathbb{R}$ a smooth objective function related to sample i or client i.

The Problem

Consider the optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{1}$$

d: the number of parameters
n: the number of samples or the number of clients participating in the learning task
$f_i : \mathbb{R}^d \to \mathbb{R}$ a smooth objective function related to sample i or client i.

The Problem

Consider the optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{1}$$

d: the number of parameters

n: the number of samples or the number of clients participating in the learning task

$f_i : \mathbb{R}^d \to \mathbb{R}$ a smooth objective function related to sample i or client i.

Assumptions

- f is bounded from below by f^*
- f is L-smooth

Zero order optimization

DFO setting

- Gradient out of the reach.
- Computing derivatives is too costly or impossible.

Zero order optimization

DFO setting

- Gradient out of the reach.
- Computing derivatives is too costly or impossible.
- Black-box objective function f: no derivative code available.

Zero order optimization

DFO setting

- Gradient out of the reach.
- Computing derivatives is too costly or impossible.
- Black-box objective function f: no derivative code available.
- Automatic differentiation: inapplicable.
  $\Rightarrow$ The gradient exists but cannot be used for algorithmic purposes.

Stochastic Three Points Method (STP)

[Bergou et al. 2020][1]

STP method

1. Choose starting iterate $x_0 \in \mathbb{R}^d$, positive stepsizes $\{\alpha_k\}_{k \geq 0}$, probability distribution $\mathcal{D}$ on $\mathbb{R}^d$.
2. **For** $k = 0, 1, 2, ...$
   1. Generate a random vector $s_k \sim \mathcal{D}$
   2. Let $x_+ = x_k + \alpha_k s_k$ and $x_- = x_k - \alpha_k s_k$
   3. $\boldsymbol{x_{k+1} = \arg\min\{f(x_-), f(x_+), f(x_k)\}}$

[1] E. Bergou, E. Gorbunov, and P. Richtárik. "Stochastic Three Points Method for Unconstrained Smooth Minimization". In: *SIAM Journal on Optimization* 30(4), 2726–2749 (2020).

Stochastic Three Points Method (STP)

[Bergou et al. 2020][1]

STP method

1. Choose starting iterate $x_0 \in \mathbb{R}^d$, positive stepsizes $\{\alpha_k\}_{k\geq 0}$, probability distribution $\mathcal{D}$ on $\mathbb{R}^d$.
2. **For** $k = 0, 1, 2, ...$
    1. Generate a random vector $s_k \sim \mathcal{D}$
    2. Let $x_+ = x_k + \alpha_k s_k$ and $x_- = x_k - \alpha_k s_k$
    3. $\boldsymbol{x_{k+1} = \arg\min\{f(x_-), f(x_+), f(x_k)\}}$

- not having exact function evaluations? (f noisy)
- evaluating f is costly? (n is too large)

[1] E. Bergou, E. Gorbunov, and P. Richtárik. "Stochastic Three Points Method for Unconstrained Smooth Minimization". In: *SIAM Journal on Optimization* 30(4), 2726–2749 (2020).

Minibatch Stochastic Three Points Method (MiSTP)

The approximation is defined as follow:

$$f_{\mathcal{B}}(x) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} f_i(x) \tag{2}$$

$\mathcal{B}$ is a randomly chosen subset of the data (or clients) and $|\mathcal{B}|$ is its cardinal.
[Boucherouite et al. 2022][2]

MiSTP method

1. **For** $k = 0, 1, 2, \ldots$
   1. Generate a random vector $s_k \sim \mathcal{D}$
   2. Choose elements of the subset $\mathcal{B}_k$
   3. Let $x_+ = x_k + \alpha_k s_k$ and $x_- = x_k - \alpha_k s_k$
   4. $\boldsymbol{x_{k+1} = \arg\min\{f_{\mathcal{B}_k}(x_-), f_{\mathcal{B}_k}(x_+), f_{\mathcal{B}_k}(x_k)\}}$

[2] S. Boucherouite et al. "Minibatch Stochastic Three Points Method for Unconstrained Smooth Minimization". In: *Submitted to ICLR* (2022).

Deterministic Direct Search

[Kolda et al. 2003]

DDS

1. Choose $x_0 \in \mathbb{R}^d$, initial stepsize $\alpha_0 > 0$, $0 < \theta < 1 < \gamma$, $c > 0$.
2. Iterate:
    1. Choose a set of directions D

Deterministic Direct Search

[Kolda et al. 2003]

DDS

1. Choose $x_0 \in \mathbb{R}^d$, initial stepsize $\alpha_0 > 0$, $0 < \theta < 1 < \gamma$, $c > 0$.
2. Iterate:
   1. Choose a set of directions D
   2. If it exists $s \in D$ s.t.

   $$f(x_k + \alpha_k s) < f(x_k) - c\alpha_k^2,$$

   then $x_{k+1} = x_k + \alpha_k s$ and $\alpha_{k+1} = \gamma \alpha_k$.

[Kolda et al. 2003]

DDS

1. Choose $x_0 \in \mathbb{R}^d$, initial stepsize $\alpha_0 > 0$, $0 < \theta < 1 < \gamma$, $c > 0$.
2. Iterate:
    1. Choose a set of directions D
    2. If it exists $s \in D$ s.t.

    $$f(x_k + \alpha_k s) < f(x_k) - c\alpha_k^2,$$

    then $x_{k+1} = x_k + \alpha_k s$ and $\alpha_{k+1} = \gamma\alpha_k$.
    3. Otherwise $x_{k+1} = x_k$ and $\alpha_{k+1} = \theta\alpha_k$.

Assumptions on D and P

The set D has the following properties

1. For $s \in D$, $\|s\|$ is positive and finite.
2. There is a constant $\mu > 0$ and norm $\| \cdot \|_P$ on $\mathbb{R}^d$ such for all $g \in \mathbb{R}^d$

$$cm(D, g) = \max_{s \in D} s^T g \geq \mu \|g\|_P$$

Assumptions on D and P

The set D has the following properties

1. For $s \in D$, $\|s\|$ is positive and finite.
2. There is a constant $\mu > 0$ and norm $\|\cdot\|_P$ on $\mathbb{R}^d$ such for all $g \in \mathbb{R}^d$

$$cm(D,g) = \max_{s \in D} s^T g \geq \mu \|g\|_P$$

The probability distribution P on $\mathbb{R}^d$ has the following properties

1. The quantity $\gamma_P \stackrel{\text{def}}{=} \mathbf{E}_{s \sim P} \|s\|_2^2$ is positive and finite.
2. There is a constant $\mu_P > 0$ and norm $\|\cdot\|_P$ on $\mathbb{R}^d$ such for all $g \in \mathbb{R}^n$,

$$\mathbf{E}_{s \sim P} \, |\, \langle g, s \rangle \,| \geq \mu_P \|g\|_P.$$

Example of D

With $g = -\nabla f(x)$, $cm(D, g) \geq \mu \|g\|_P$ means that D contains a descent direction for f at x.

Example of D

1. $D = \{e_1, \ldots, e_d, -e_1, \ldots, -e_d\}$
2. $cm(D, g) \geq \frac{1}{\sqrt{d}} \|g\|_2$

$$cm(D, g) = \max_{d \in D} d^T g \geq \mu \|g\|_P$$

Example of D

With $g = -\nabla f(x)$, $cm(D, g) \geq \mu \|g\|_P$ means that D contains a descent direction for f at x.

Example of D

1. $D = \{e_1, \ldots, e_d, -e_1, \ldots, -e_d\}$
2. $cm(D, g) \geq \frac{1}{\sqrt{d}} \|g\|_2$

$$cm(D, g) = \max_{d \in D} d^T g \geq \mu \|g\|_P$$

3. In general D must contains at least $d + 1$ vectors.

Examples of P, continuous distributions

1. If P is the uniform distribution on the unit sphere in $\mathbb{R}^d$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P}\,|\,\langle g, s\rangle\,| \sim \frac{1}{\sqrt{2\pi d}}\|g\|_2.$$

Examples of P, continuous distributions

1. If P is the uniform distribution on the unit sphere in $\mathbb{R}^d$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P} \,|\, \langle g, s\rangle \,|\, \sim \frac{1}{\sqrt{2\pi d}}\|g\|_2.$$

2. If P is the normal distribution with zero mean and identity over d as covariance matrix i.e. $s \sim N(0, \frac{I}{d})$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P} \,|\, \langle g, s\rangle \,|\, = \frac{\sqrt{2}}{\sqrt{d\pi}}\|g\|_2.$$

Examples of P, discrete distributions

1. If P is the uniform distribution on $\{e_1, \ldots, e_d\}$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P} \mid \langle g, s\rangle \mid = \frac{1}{d}\|g\|_1.$$

Examples of P, discrete distributions

1. If P is the uniform distribution on $\{e_1, \ldots, e_d\}$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P} \mid \langle g, s\rangle \mid = \frac{1}{d}\|g\|_1.$$

2. If P is an arbitrary distribution on $\{e_1, \ldots, e_d\}$ given by $P(s = e_i) = p_i > 0$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P} \mid \langle g, s\rangle \mid = \|g\|_p \overset{\text{def}}{=} \sum_{i=1}^{d} p_i|g_i|.$$

Examples of P, discrete distributions

1. If P is the uniform distribution on $\{e_1, \ldots, e_d\}$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P} \mid \langle g, s\rangle \mid = \frac{1}{d}\|g\|_1.$$

2. If P is an arbitrary distribution on $\{e_1, \ldots, e_d\}$ given by $P(s = e_i) = p_i > 0$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P} \mid \langle g, s\rangle \mid = \|g\|_p \overset{\text{def}}{=} \sum_{i=1}^{d} p_i|g_i|.$$

3. If P is a distribution on $D = \{s_1, \ldots, s_d\}$ where $s_1, \ldots, s_d$ form an orthonormal basis of $\mathbb{R}^n$ and $P(s = s_i) = p_i$, then

$$\gamma_P = \mathbf{E}_{s\sim P}\|s\|_2^2 = 1 \text{ and } \mathbf{E}_{s\sim P} \mid \langle g, s\rangle \mid = \|g\|_p \overset{\text{def}}{=} \sum_{i=1}^{d} p_i|g_i|.$$

Worst-case complexity of DDS

[Vicente 2013]

Evaluation complexity

Let $\epsilon \in (0,1)$ and $K(\epsilon)$ be the number of function evaluations needed to reach a point such that

$$\min_{k=0,1,\ldots,K} [\|\nabla f(x_k)\|_P] \leq \epsilon.$$

Then

$$K(\epsilon) \leq O(|D|(\mu\epsilon)^{-2}).$$

Worst-case complexity of DDS

[Vicente 2013]

Evaluation complexity

Let $\epsilon \in (0,1)$ and $K(\epsilon)$ be the number of function evaluations needed to reach a point such that

$$\min_{k=0,1,\dots,K} \left[\|\nabla f(x_k)\|_P\right] \leq \epsilon.$$

Then

$$K(\epsilon) \leq O(|D|(\mu\epsilon)^{-2}).$$

With $D = \{e_1, \dots, e_d, -e_1, \dots, -e_d\}$, we have $|D| = 2n$ and $\mu = \frac{1}{\sqrt{d}}$ thus

$$K(\epsilon) \leq O(d^2 \epsilon^{-2}).$$

Worst-case complexity of STP and MiSTP

Evaluation complexity

Let $\epsilon \in (0,1)$ and $K(\epsilon)$ be the number of function evaluations needed to reach a point such that

$$\min_{k=0,1,\ldots,K} \mathbf{E}\left[\|\nabla f(x_k)\|_P\right] \leq \epsilon.$$

Then

$$K(\epsilon) \leq O((\mu\epsilon)^{-2}).$$

Worst-case complexity of STP and MiSTP

Evaluation complexity

Let $\epsilon \in (0, 1)$ and $K(\epsilon)$ be the number of function evaluations needed to reach a point such that

$$\min_{k=0,1,\ldots,K} \mathbf{E}\left[\|\nabla f(x_k)\|_P\right] \le \epsilon.$$

Then

$$K(\epsilon) \le O((\mu\epsilon)^{-2}).$$

$\mu \sim \frac{cst}{\sqrt{d}}$ thus

$$K(\epsilon) \le O(d\epsilon^{-2}).$$

Worst-case complexity of DDS under convexity

[Vicente et al. 2016]

Evaluation complexity

Let $\epsilon \in (0, 1)$ and $K(\epsilon)$ be the number of function evaluations needed to reach a point such that

$$f(x_k) - f(x^*) \leq \epsilon.$$

Then $K(\epsilon) \leq O(|D|\mu^{-2}\epsilon^{-1})$.

Worst-case complexity of DDS under convexity

[Vicente et al. 2016]

Evaluation complexity

Let $\epsilon \in (0,1)$ and $K(\epsilon)$ be the number of function evaluations needed to reach a point such that

$$f(x_k) - f(x^*) \leq \epsilon.$$

Then $K(\epsilon) \leq O(|D|\mu^{-2}\epsilon^{-1})$.

$|D| = 2d$ and $\mu = \frac{1}{\sqrt{d}}$ thus $K(\epsilon) \leq O(d^2\epsilon^{-1})$.

Worst-case complexity of STP and MiSTP under convexity

Evaluation complexity

Let $\epsilon \in (0,1)$ and $K(\epsilon)$ be the number of function evaluations needed to reach a point such that

$$\mathbf{E}\left[f(x_k) - f(x^*)\right] \leq \epsilon.$$

Then $K(\epsilon) \leq O(\mu^{-2}\epsilon^{-1})$.

Worst-case complexity of STP and MiSTP under convexity

Evaluation complexity

Let $\epsilon \in (0,1)$ and $K(\epsilon)$ be the number of function evaluations needed to reach a point such that

$$\mathbf{E}\left[f(x_k) - f(x^*)\right] \leq \epsilon.$$

Then $K(\epsilon) \leq O(\mu^{-2}\epsilon^{-1})$.

$\mu \sim \frac{cst}{\sqrt{d}}$ thus $K(\epsilon) \leq O(d\epsilon^{-1})$.

(Mi) STP vs DDS

The main differences between DDS and (Mi)STP are:

- (Mi)STP uses one random direction at each iteration while DDS uses many deterministic directions (at least $d+1$).

(Mi) STP vs DDS

The main differences between DDS and (Mi)STP are:

- (Mi)STP uses one random direction at each iteration while DDS uses many deterministic directions (at least $d + 1$).
- DDS imposes sufficient decrease condition to accept the iterates.

(Mi) STP vs DDS

The main differences between DDS and (Mi)STP are:

- (Mi)STP uses one random direction at each iteration while DDS uses many deterministic directions (at least $d + 1$).
- DDS imposes sufficient decrease condition to accept the iterates.
- DDS updates step size automatically while in (Mi)STP one needs to choose the step sizes at the beginning of the algorithm.

(Mi) STP vs DDS

The main differences between DDS and (Mi)STP are:

- (Mi)STP uses one random direction at each iteration while DDS uses many deterministic directions (at least $d + 1$).
- DDS imposes sufficient decrease condition to accept the iterates.
- DDS updates step size automatically while in (Mi)STP one needs to choose the step sizes at the beginning of the algorithm.
- In (Mi)STP many choices of the step sizes apply.

(Mi) STP vs DDS

The main differences between DDS and (Mi)STP are:

- (Mi)STP uses one random direction at each iteration while DDS uses many deterministic directions (at least $d + 1$).
- DDS imposes sufficient decrease condition to accept the iterates.
- DDS updates step size automatically while in (Mi)STP one needs to choose the step sizes at the beginning of the algorithm.
- In (Mi)STP many choices of the step sizes apply.
- The complexity of (Mi)STP depends linearly in d while its dependence is quadratic in DDS.

- Ridge regression:

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n}(A[i,:]x - y_i)^2 + \frac{\lambda}{2}\|x\|_2^2$$

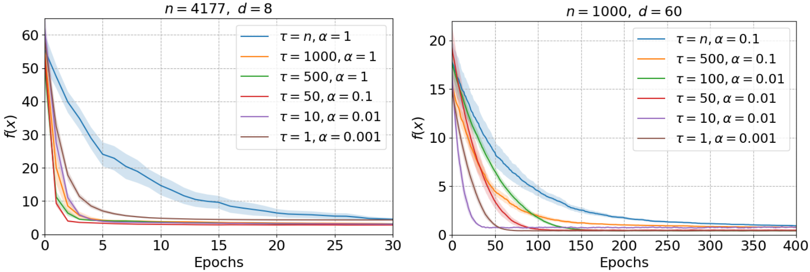

Figure 1: Performance of `MiSTP` with different minibatch sizes on ridge regression problem. On the left, the "abalone" dataset. On the right, the "splice" dataset from LIBSVM.

MiSTP on the logistic regression problem

- Regularized logistic regression:

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n} \ln(1+\exp(-y_i A[i,:]x)) + \frac{\lambda}{2}\|x\|_2^2$$

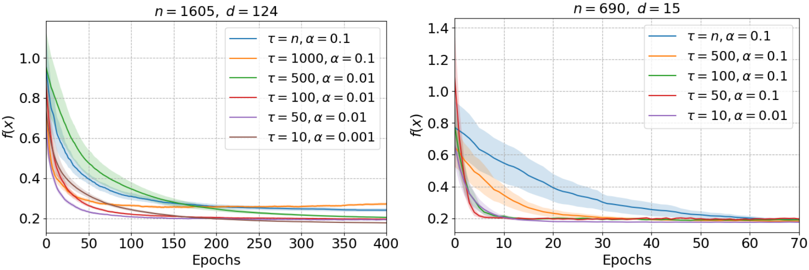

Figure 2: Performance of `MiSTP` with different minibatch sizes on regularized logistic regression problem. On the left, the "a1a" dataset. On the right, the "australian" dataset.

MiSTP vs. SGD

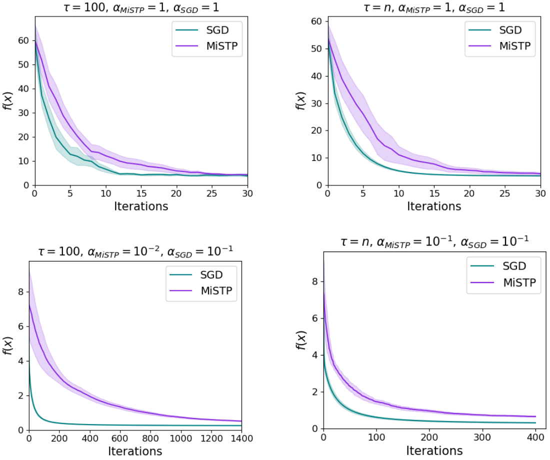

Figure 3: Performance of `MiSTP` and `SGD` on ridge regression problem using real data from LIBSVM. Above, abalone dataset: $n = 4177$ and $d = 8$. Below, a1a dataset :$n = 1605$ and $d = 123$.

MiSTP vs. SGD

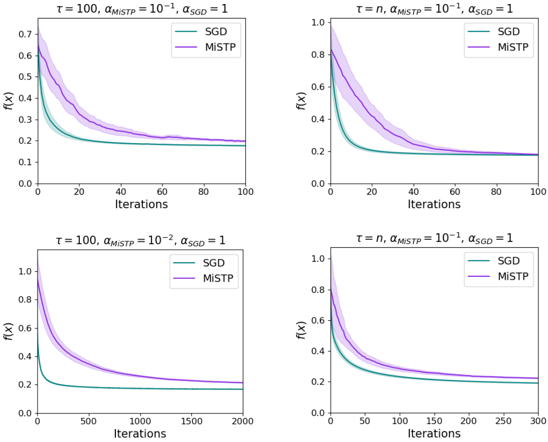

Figure 4: Performance of `MiSTP` and `SGD` on regularized logistic regression problem using real data from LIBSVM. Above, australian dataset : $n = 690$ and $d = 15$. Below, a1a dataset : $n = 1605$ and $d = 124$.

MiSTP vs. other zero order methods

- `RSGF` (Random Stochastic Gradient Free)[3] :

$$x_{k+1} = x_k - \alpha_k \frac{f_{\mathcal{B}_k}(x_k + \mu_k s_k) - f_{\mathcal{B}_k}(x_k)}{\mu_k} s_k$$

- `ZO-SVRG`(Zero Order Stochastic variance reduced Gradient)[4]:

$$\hat{\nabla} f_{\mathcal{B}_k}(x_k) = \frac{d}{\mu}(f_{\mathcal{B}_k}(x_k + \mu s_k) - f_{\mathcal{B}_k}(x_k)) s_k$$

- `ZO-CD` (Zero Order Coordinate descent):

$$x_{k+1} = x_k - \alpha_k g_{\mathcal{B}_k}, \quad g_{\mathcal{B}_k} = \sum_{i=1}^{d} \frac{f_{\mathcal{B}_k}(x_k + \mu e_i) - f_{\mathcal{B}_k}(x_k - \mu e_i)}{2\mu} e_i$$

---

[3] S. Ghadimi and G. Lan. "Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming". In: *SIAM Journal on Optimization* 23.4 (2013), pp. 2341–2368.

[4] S. Liu et al. "Zeroth-order stochastic variance reduction for nonconvex optimization". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2018), pp. 3731–3741.

MiSTP vs. other zero order methods

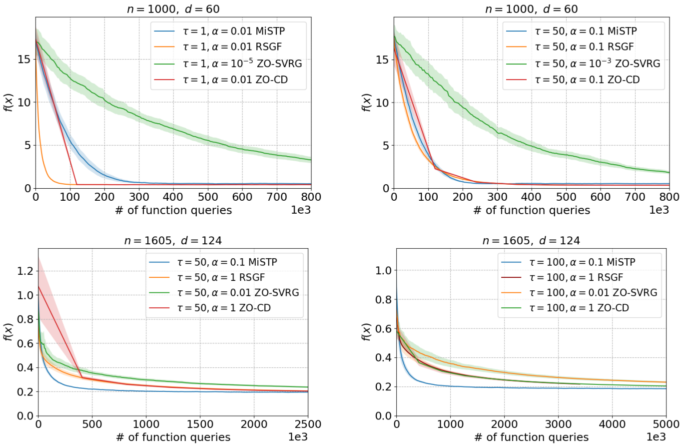

Figure 5: Comparison of `MiSTP`, `RSGF`, `ZO-SVRG`, and `ZO-CD`. Above: ridge regression problem using the "splice" dataset. Below: regularized logistic regression problem using the "a1a" dataset.

- MNIST digit classification
- Three fully-connected layers of size 256, 128, 10, with ReLU activation after the first two layers and a Softmax activation function after the last layer.
- The loss function: the categorical cross entropy.

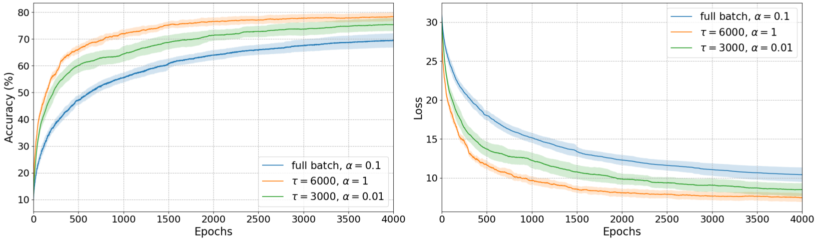

Figure 6: Comparison of different minibatch sizes for `MiSTP` in a multi-layer neural network.

Conclusion & Perspectives

Conclusions

- Simple STP approach for DFO.

Conclusion & Perspectives

Conclusions

- Simple STP approach for DFO.
- The worst case complexity of (Mi)STP depends linearly on d, and the same way as the steepest descent on ϵ.

Conclusion & Perspectives

Conclusions

- Simple STP approach for DFO.
- The worst case complexity of (Mi)STP depends linearly on d, and the same way as the steepest descent on ϵ.

Some perspectives

- Parallel version of STP.

Conclusion & Perspectives

Conclusions

- Simple STP approach for DFO.
- The worst case complexity of (Mi)STP depends linearly on d, and the same way as the steepest descent on ϵ.

Some perspectives

- Parallel version of STP.
- Extension to the constrained optimization.

Conclusion & Perspectives

Conclusions

- Simple STP approach for DFO.
- The worst case complexity of (Mi)STP depends linearly on d, and the same way as the steepest descent on ϵ.

Some perspectives

- Parallel version of STP.
- Extension to the constrained optimization.
- Deriving a rule to find the optimal minibatch size.

Conclusion & Perspectives

Conclusions

- Simple STP approach for DFO.
- The worst case complexity of (Mi)STP depends linearly on d, and the same way as the steepest descent on ϵ.

Some perspectives

- Parallel version of STP.
- Extension to the constrained optimization.
- Deriving a rule to find the optimal minibatch size.
- Investigating `MiSTP` in the non-smooth case.
- $\dots$

Thank you for your attention!