Adaptive primal-dual stochastic gradient methods

Yangyang Xu

Mathematical Sciences, Rensselaer Polytechnic Institute

October 26, 2019

Stochastic gradient method

stochastic program:

$$\min_{\mathbf{x}\in X} f(\mathbf{x}) = \mathbb{E}_{\xi}\left[F(\mathbf{x};\xi)\right]$$

- if ξ uniform on $\{\xi_1,\ldots,\xi_N\}$, then $f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} F(\mathbf{x};\xi_i)$
- stochastic gradient (that requires samples of ξ):

$$\mathbf{x}^{k+1} = \mathrm{Proj}_X\left(\mathbf{x}^k - \alpha_k \mathbf{g}_k\right)$$

  where $\mathbf{g}_k$ is a stochastic approximation of $\nabla f(\mathbf{x}^k)$
- low per-update complexity compared to deterministic gradient descent
- Literature: tons of works (e.g., [Robbins-Monro'51, Polyak-Juditsky'92, Nemirovski et. al. '09, Ghadimi-Lan'13, Davis et. al'18])

adaptive learning

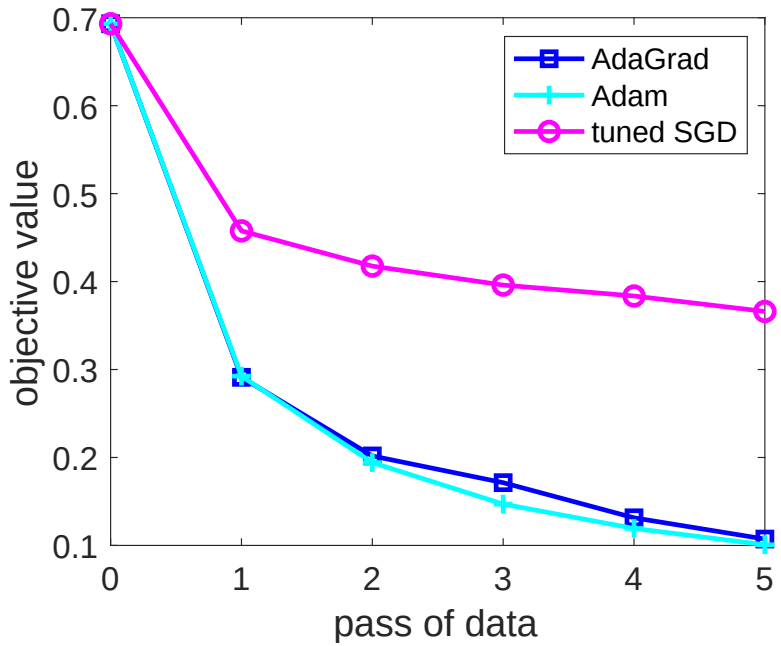
- adaptive gradient [Duchi-Hazan-Singer'11]:

$$\mathbf{x}^{k+1} = \mathrm{Proj}_X^{\mathbf{v}_k}\left(\mathbf{x}^k - \alpha_k \cdot \mathbf{g}_k \oslash \mathbf{v}_k\right)$$

  where $\mathbf{v}_k = \sqrt{\sum_{t=0}^{k} (\mathbf{g}_k)^2}$

- many other adaptive variants: Adam [Kingma-Ba'14], AMSGrad [Reddi-Kale-Kumar'18], and so on
- extremely popular in training deep neural networks

Adaptiveness improves convergence speed



- test on solving a neural network with one hidden layer

Observation: adaptive methods much faster, and all methods have similar per-update cost

Take a close look: $\mathbf{x}^{k+1} = \mathrm{Proj}_X^{\mathbf{v}_k}\left(\mathbf{x}^k - \alpha_k \cdot \mathbf{g}_k \oslash \mathbf{v}_k\right)$

- $\mathrm{Proj}_X^{\mathbf{v}_k}$ is assumed simple (holds if X is simple)
- Not (easily) implementable if X is complicated

This talk: adaptive primal-dual stochastic gradient for problems with complicated constraints

Outline

1. Problem formula and motivating examples

2. Review of existing methods

3. Proposed primal-dual stochastic gradient method

4. Numerical and convergence results and conclusions

Stochastic functional constrained stochastic program

$$\begin{aligned} &\min_{\mathbf{x}\in X} \; f_0(\mathbf{x}) = \mathbb{E}_{\xi_0}\left[F_0(\mathbf{x};\xi_0)\right], \\ &\text{s.t.} \;\; f_j(\mathbf{x}) = \mathbb{E}_{\xi_j}\left[F_j(\mathbf{x};\xi_j)\right] \leq 0,\, j = 1,\ldots,m \end{aligned} \tag{P}$$

- X is a simple closed convex set (but the feasible set is complicated)
- f_j is convex and possibly nondifferentiable
- m could be very big: expensive to access all f_j's at every update

Goal: design an efficient stochastic method without complicated projection that can guarantees (near) optimality and feasibility

Example I: linear programming of Markov decision process

discounted Markov decision process: $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \gamma)$

- state space $\mathcal{S} = \{s_1, \ldots, s_m\}$, action space $\mathcal{A} = \{a_1, \ldots, a_n\}$
- transition probability $\mathcal{P} = [P_a(s, s')]$, reward $\mathbf{r} = [r_a(s, s')]$
- discount factor: $\gamma \in (0, 1]$

Bellman optimality equation:

$$v(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_a(s, s') \left[r_a(s, s') + \gamma v(s') \right], \forall s \in \mathcal{S}$$

equivalent to linear programming [Puterman'14]:

$$\min_{\mathbf{v}} \mathbf{e}^\top \mathbf{v}, \text{ s.t. } (\mathbf{I} - \gamma \mathbf{P}_a)\mathbf{v} - \mathbf{r}_a \geq \mathbf{0}, \forall a \in \mathcal{A}$$

- $r_a(s) = \sum_{s' \in \mathcal{S}} P_a(s, s') r_a(s, s')$
- huge number of constraints if m and/or n is big

Example II: robust optimization by sampling

Robust optimization:

$$\min_{\mathbf{x}\in X} f_0(\mathbf{x}),\ \text{s.t. } g(\mathbf{x};\xi) \leq 0,\ \forall \xi \in \Xi$$

Sampled approximation [Calafiore-Campi'05]:

$$\min_{\mathbf{x}\in X} f_0(\mathbf{x}),\ \text{s.t. } g(\mathbf{x};\xi_i) \leq 0, \forall i = 1,\ldots,m$$

- $\{\xi_1,\ldots,\xi_m\}$: m independently extracted samples
- solution of the sampled approximation problem is a $(1-\tau)$-level robustly feasible solution with probability at least $1-\varepsilon$ if

$$m \geq \frac{n}{\tau\varepsilon} - 1,$$

where $\tau \in (0,1)$ and $\varepsilon \in (0,1)$.

Literature

Few for problems with functional constraints

- penalty method with stochastic approximation [Wang-Ma-Yuan'17]
  - uses exact function/gradient information of all constraint functions
- stochastic mirror-prox descent for saddle-point problems [Baes-Brgisser-Nemirovski'13]
- cooperative stochastic approximation (CSA) for problems with expectation constraint [Lan-Zhou'16]
- level-set methods [Lin et. al'18]

Stochastic mirror-prox method [Baes-Brgisser-Nemirovski'13]

For a saddle-point problem:

$$\min_{\mathbf{x}\in X}\max_{\mathbf{z}\in Z}\mathcal{L}(\mathbf{x},\mathbf{z})$$

Iterative update scheme:

$$\left(\hat{\mathbf{x}}^k,\hat{\mathbf{z}}^k\right)=\mathrm{Proj}_{X\times Z}\left(\left(\mathbf{x}^k-\alpha_k\mathbf{g}_x^k,\mathbf{z}^k+\alpha_k\mathbf{g}_z^k\right)\right),$$

$$\left(\mathbf{x}^{k+1},\mathbf{z}^{k+1}\right)=\mathrm{Proj}_{X\times Z}\left(\left(\mathbf{x}^k-\alpha_k\hat{\mathbf{g}}_x^k,\mathbf{z}^k+\alpha_k\hat{\mathbf{g}}_z^k\right)\right)$$

- $(\mathbf{g}_x^k;\mathbf{g}_z^k)$: a stochastic approximation of $\nabla\mathcal{L}(\mathbf{x}^k,\mathbf{z}^k)$
- $(\hat{\mathbf{g}}_x^k;\hat{\mathbf{g}}_z^k)$: a stochastic approximation of $\nabla\mathcal{L}(\hat{\mathbf{x}}^k,\hat{\mathbf{z}}^k)$
- $O(1/\sqrt{k})$ rate in terms of primal-dual gap

Cooperative stochastic approximation [Lan-Zhou'16]

For the problem with expectation constraint:

$$\min_{\mathbf{x}\in X} f(\mathbf{x}) = \mathbb{E}_\xi[F(\mathbf{x},\xi)],\ \text{s.t. } \mathbb{E}_\xi[G(\mathbf{x},\xi)] \le 0$$

For $k = 0, 1, \ldots,$ do

1. sample ξ_k;
2. If $G(\mathbf{x}^k, \xi_k) \le \eta_k$, set $\mathbf{g}^k = \tilde{\nabla} F(\mathbf{x}^k, \xi_k)$; otherwise, $\mathbf{g}^k = \tilde{\nabla} G(\mathbf{x}^k, \xi_k)$
3. Update $\mathbf{x}$ by
$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in X} \langle \mathbf{g}^k, \mathbf{x}\rangle + \frac{1}{2\gamma_k}\|\mathbf{x} - \mathbf{x}^k\|^2$$

- purely primal method
- $O(1/\sqrt{k})$ rate for convex problems
- $O(1/k)$ if both objective and constraint functions are strongly convex

proposed method by the augmented Lagrangian function

Augmented Lagrangian function

With slack variables $\mathbf{s} \geq \mathbf{0}$, (P) is equivalent to

$$\min_{\mathbf{x}\in X, \mathbf{s}\geq \mathbf{0}} f_0(\mathbf{x}), \text{ s.t. } f_i(\mathbf{x}) + s_i = 0, i = 1, \ldots, m.$$

By quadratic penalty, the augmented Lagrangian function is

$$\tilde{\mathcal{L}}_\beta(\mathbf{x}, \mathbf{s}, \mathbf{z}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} z_i \big(f_i(\mathbf{x}) + s_i\big) + \frac{\beta}{2} \sum_{i=1}^{m} \big(f_i(\mathbf{x}) + s_i\big)^2.$$

Fix $(\mathbf{x}, \mathbf{z})$ and minimize $\tilde{\mathcal{L}}_\beta$ about $\mathbf{s} \geq \mathbf{0}$ (through solving $\nabla_{\mathbf{s}} \tilde{\mathcal{L}}_\beta = \mathbf{0}$):

$$s_i = \left[-\frac{z_i}{\beta} - f_i(x)\right]_+, \ i = 1, \ldots, m.$$

Augmented Lagrangian function

Eliminate $\mathbf{s}$ to have the classic augmented Lagrangian function of (P):

$$\mathcal{L}_\beta(\mathbf{x}, \mathbf{z}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \psi_\beta(f_i(\mathbf{x}), z_i),$$

where

$$\psi_\beta(u, v) = \begin{cases} uv + \frac{\beta}{2}u^2, & \text{if } \beta u + v \geq 0, \\ -\frac{v^2}{2\beta}, & \text{if } \beta u + v < 0. \end{cases}$$

- $\psi_\beta(f_i(\mathbf{x}), z_i)$ convex in $\mathbf{x}$ and concave in z_i for each i
- thus $\mathcal{L}_\beta$ convex in $\mathbf{x}$ and concave in $\mathbf{z}$

Augmented Lagrangian method

Choose $(\mathbf{x}^1, \mathbf{y}^1, \mathbf{z}^1)$. For $k = 1, 2, \ldots$, iteratively do:

$$\mathbf{x}^{k+1} \in \operatorname*{Arg\,min}_{\mathbf{x} \in X} \mathcal{L}_\beta(\mathbf{x}, \mathbf{z}^k),$$
$$\mathbf{z}^{k+1} = \mathbf{z}^k + \rho \nabla_{\mathbf{z}} \mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{z}^k)$$

- if $\rho < 2\beta$, globally convergent with rate $O\left(\frac{1}{k\rho}\right)$
- bigger ρ and β gives faster convergence in term of iteration number but yields harder $\mathbf{x}$-subproblem

Proposed primal-dual stochastic gradient method

Consider the case:

- exact f_j and $\tilde{\nabla} f_j$ can be obtained for each $j = 1, \ldots, m$
- m is big: expensive to access all f_j's every update

Examples: MDP, robust optimization by sampling, multi-class SVM

Remarks:

- if f_j is stochastic, AL function is a compositional expectation form
  - difficult to obtain unbiased stochastic estimation of $\tilde{\nabla}_{\mathbf{x}} \mathcal{L}_\beta$
- ordinary Lagrangian function can be used to handle the most general case

Proposed primal-dual stochastic gradient method

For $k = 0, 1, \ldots,$ do

1. Sample ξ_k and pick $j_k \in [m]$ uniformly at random;
2. Let $\mathbf{g}^k = \tilde{\nabla} F_0(\mathbf{x}^k, \xi_k) + \tilde{\nabla}_{\mathbf{x}} \psi_\beta \left(f_{j_k}(\mathbf{x}^k), z_{j_k}^k \right)$;
3. Update the primal variable $\mathbf{x}$ by
$$\mathbf{x}^{k+1} = \mathrm{Proj}_X \left(\mathbf{x}^k - \mathbf{D}_k^{-1} \mathbf{g}^k \right)$$
4. Let $z_j^{k+1} = z_j^k$ for $j \neq j_k$ and update z_{j_k} by
$$z_j^{k+1} = z_j^k + \rho_k \cdot \max \left(-\frac{z_j^k}{\beta}, f_j(\mathbf{x}^k) \right), \text{ for } j = j_k.$$

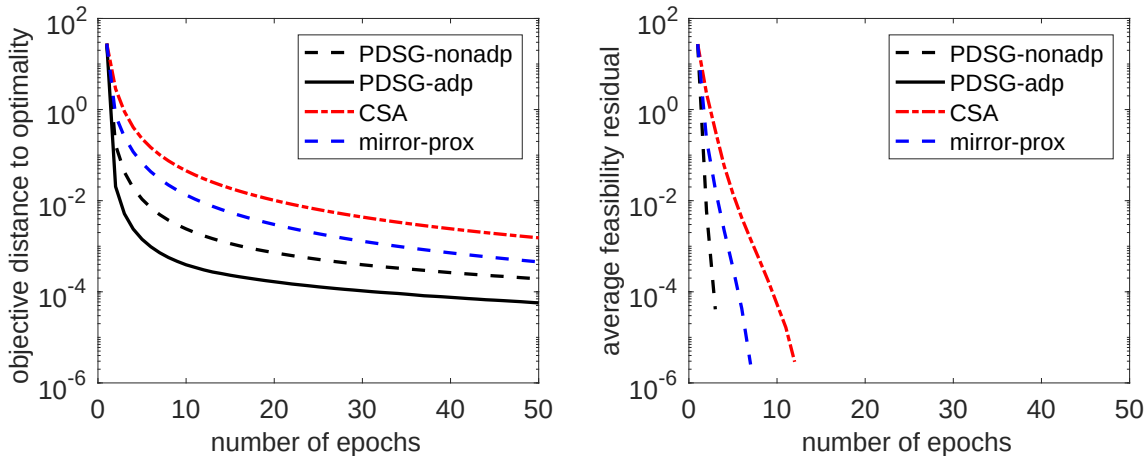
- $\mathbf{g}^k$ unbiased stochastic estimation of $\tilde{\nabla}_{\mathbf{x}} \mathcal{L}_\beta$ at $\mathbf{x}^k$
- $\tilde{\nabla} f_{j_k}(\mathbf{x}^k)$ required, and $f_{j_k}(\mathbf{x}^k)$ and $f_{j_k}(\mathbf{x}^k)$ needed for the updates
- $\mathbf{D}_k = \mathbf{I}/\alpha_k + \eta \cdot \mathrm{diag} \left(\sqrt{\sum_{t=0}^k |\tilde{\mathbf{g}}^k|^2} \right)$ with $\tilde{\mathbf{g}}^k$ scaled version of $\mathbf{g}^k$

How the proposed method performs

Test on convex quadratically constrained quadratic programming

$$\min_{\mathbf{x}\in X} \frac{1}{2N}\sum_{i=1}^{N}\|\mathbf{H}_i\mathbf{x}-\mathbf{c}_i\|^2,\ \text{s.t.}\ \frac{1}{2}\mathbf{x}^\top\mathbf{Q}_j\mathbf{x}+\mathbf{a}_j^\top\mathbf{x}\le b_j, j=1,\dots,m,$$

where $N = m = 10,000$.



Observations:

- proposed methods better than mirror-prox and CSA
- adaptiveness significantly improves convergence speed
- all methods have roughly the same asymptotic convergence rate

Sublinear convergence result

Assumptions:

1. existence of a primal-dual solution $(\mathbf{x}^*, \mathbf{z}^*)$
2. unbiased estimate and bounded variance
3. bounded constraint function and subgradient

Theorem: Given K, let $\alpha_k = \frac{\alpha}{\sqrt{K}}, \rho_k = \frac{\rho}{\sqrt{K}}, \beta \geq \rho$. Then

$$\max\left(\mathbb{E}\left|f_0(\bar{\mathbf{x}}^K) - f_0(\mathbf{x}^*)\right|, \mathbb{E}\|[\mathbf{f}(\bar{\mathbf{x}}^K)]_+\|\right) = O\left(1/\sqrt{K}\right)$$

If f_0 is strongly convex, let $\alpha_k = \frac{\alpha}{k}, \rho_k = \frac{\rho}{\log(K+1)}, \beta \geq \frac{2\rho}{\log 2}$. Then

$$\mathbb{E}\|\mathbf{x}^K - \mathbf{x}^*\|^2 = O\left(\frac{\log(K+1)}{K}\right)$$

- $\bar{\mathbf{x}}^K$ weighted average of $\{\mathbf{x}^k\}_{k=1}^{K+1}$

Remark: CSA [Lan-Zhou'16] requires strong convexity of both objective and constraint functions to achieve $O(\frac{1}{K})$

Conclusions

- Proposed an adaptive primal-dual stochastic gradient method for stochastic programs with many functional constraints
  - Based on the classic augmented Lagrangian function
  - $O(1/\sqrt{k})$ convergence for convex problems
  - $O\big((\log k)/k\big)$ convergence if the objective is strongly convex
- Numerical experiment on a convex quadratically constrained quadratic program
  - better than two state-of-the-art methods
  - adaptiveness can significantly improve the convergence speed

References

Y. Xu. Primal-dual stochastic gradient method for convex programs with many functional constraints, arXiv:1802.02724, 2018.

Thank you!!!