Universal Stochastic Gradient Methods for Convex Optimization

Anton Rodomanov (UCLouvain)

Jan 31, 2023

CISPA Helmholtz Center for Information Security, Saarbrücken

Part I: Motivation

Stochastic Gradient Method (SGD)

Problem:

$$f^* = \min_{x \in Q} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $Q \subseteq \mathbb{R}^n$ is a simple convex set.

Stochastic gradient oracle: Vector $g(x, \xi) \in \mathbb{R}^n$, $\xi \sim P_\xi$ with

$$\mathbb{E}_\xi[g(x, \xi)] = \nabla f(x).$$

SGD algorithm:

$$x_{k+1} = \pi_Q(x_k - h_k g_k), \quad g_k = g(x_k, \xi_k), \quad \xi_k \sim P_\xi,$$

where $\pi_Q(x) = \operatorname{argmin}_{y \in Q} \|x - y\|$ is the Euclidean projection onto Q.

Main question: How to choose step sizes h_k?

Choice of Step Size in SGD

Assume that Q is bounded: $\|x - y\| \leq D$, $\forall x, y \in Q$.

Nonsmooth optimization: $\mathbb{E}_\xi[\|g(x,\xi)\|^2] \leq M^2$.

$$h_k = \frac{D}{M\sqrt{k+1}} \quad \Longrightarrow \quad \mathbb{E}[f(\bar{x}_k)] - f^* \leq O\Big(\frac{MD}{\sqrt{k}}\Big),$$

where $\bar{x}_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i$.

Smooth optimization: $\begin{cases} \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \\ \mathbb{E}_\xi[\|g(x,\xi) - \nabla f(x)\|^2] \leq \sigma^2. \end{cases}$

$$h_k = \min\Big\{\frac{1}{2L}, \frac{D}{\sigma\sqrt{k+1}}\Big\} \quad \Longrightarrow \quad \mathbb{E}[f(\bar{x}_k)] - f^* \leq O\Big(\frac{LD^2}{k} + \frac{\sigma D}{\sqrt{k}}\Big).$$

Can the algorithm choose step sizes automatically for us?

Line Search for Deterministic Optimization

Universal Gradient Method (UGM): [Nesterov'14]

Input: Initial point $x_0 \in Q$, target accuracy $\epsilon > 0$, initial guess $\tilde{L}_0 > 0$.

for $k \geq 0$ **do**

Set $L_{k,0} = \tilde{L}_k$.

for $i \geq 0$ **do**

Compute $x_{k+1,i} = \pi_Q\big(x_k - \frac{1}{L_{k,i}} \nabla f(x_k)\big)$.

if

$$f(x_{k+1,i}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1,i} - x_k \rangle + \frac{L_{k,i}}{2} \|x_{k+1,i} - x_k\|^2 + \frac{\epsilon}{2},$$

then

set $i_k = i$ and break the loop.

Set $L_{k,i+1} = 2L_{k,i}$.

Set $x_{k+1} = x_{k+1,i_k}$, $L_k = L_{k,i_k}$ and $\tilde{L}_{k+1} = L_k/2$.

Convergence Rate

Hölder class: $\|\nabla f(x) - \nabla f(y)\| \leq L_\nu \|x - y\|^\nu$, $\nu \in [0, 1]$.

- For properly chosen $\tilde{L}_0$ and the "best point" x_k^* in UGM:

$$k \geq \left(\frac{L_\nu}{\epsilon}\right)^{2/(1+\nu)} D^2 \quad \Longrightarrow \quad f(x_k^*) - f^* \leq \epsilon.$$

- Line search is cheap: two iterations on average (+ log-cost warmup).
- Universal Fast Gradient Method [Nesterov'14]:

$$k \geq \left(\frac{L_\nu D^{1+\nu}}{\epsilon}\right)^{2/(1+3\nu)} \quad \Longrightarrow \quad f(x_k) - f^* \leq \epsilon.$$

AdaGrad Method [Duchi et al.'11]

AdaGrad algorithm:

$$x_{k+1} = \pi_Q(x_k - h_k g_k), \qquad h_k = \frac{D}{\sqrt{\sum_{i=0}^{k} \|g_i\|^2}}.$$

Foundation of nowadays popular Adam, RMSProp, ...

Convergence rate:

$$\mathbb{E}[f(\bar{x}_k)] - f^* \leq \frac{MD}{\sqrt{k}}.$$

Main issues:

- Smooth optimization?
- Acceleration?

Recent Work

UniXGrad: [Kavis et al.'19]

Input: Initial point $y_0 \in Q$, diameter D.

Set $\bar{x}_0 = y_0$, $A_0 = 0$

for $k \geq 1$ **do**

Set $a_k = k$, $A_k = A_{k-1} + a_k$, $\tau_k = a_k / A_k$.

Compute $\tilde{z}_k = (1 - \tau_k)\bar{x}_k + \tau_k y_{k-1}$, $g_k^{\tilde{z}} = g(\tilde{z}_k, \xi_k^{\tilde{z}})$ for $\xi_k^{\tilde{z}} \sim P_\xi$.

Set $x_k = \pi_Q(y_{k-1} - a_k \eta_k g_k^{\tilde{z}})$.

Compute $\bar{x}_k = (1 - \tau_k)\bar{x}_{k-1} + \tau_k x_k$, $g_k^{\bar{x}} = g(\bar{x}_k, \xi_k^{\bar{x}})$ for $\xi_k^{\bar{x}} \sim P_\xi$.

Set $y_k = \pi_Q(y_{k-1} - a_k \eta_k g_k^{\bar{x}})$.

Step size:

$$\eta_k = \frac{2D}{\sqrt{1 + \sum_{i=1}^{k-1} a_i^2 \|g_i^{\bar{x}} - g_i^{\tilde{z}}\|^2}}.$$

Recent Work (cont'd)

Convergence rates for UniXGrad:

Nonsmooth case: $\mathbb{E}[f(\bar{x}_k)] - f^* \le O\Big(\frac{D}{k^2} + \frac{MD}{\sqrt{k}}\Big)$,

Smooth case: $\mathbb{E}[f(\bar{x}_k)] - f^* \le O\Big(\frac{LD^2}{k^2} + \frac{\sigma D}{\sqrt{k}}\Big)$.

Note: "Incorrect" rate for nonsmooth case (wrong physical dimension) $\Longleftarrow$ bad choice of η_k (not scale-invariant).

Another line of work: [Ene et al.'21]

- AdaGrad+ and AdaACSA with rates for smooth / nonsmooth cases.
- Also non-scale-invariant step size + strange log factors.

Motivation for This Work

Minor: Propose improved versions of UniXGrad, etc. with scale-invariant step sizes and "correct" convergence rates.

Major: Develop "fully universal" stochastic gradient methods with guarantees for the entire Hölder class (c.f. deterministic methods with line search).

$\Longrightarrow$ More efficient practical algorithms?

Part II: Our Results

Problem Formulation

Composite optimization problem:

$$F^* = \min_{x \in \operatorname{dom} \psi} \{F(x) = f(x) + \psi(x)\},$$

where f and ψ are convex functions, ψ is simple.

Assumptions:

1. Bounded domain: $\|x - y\| \leq D, \quad \forall x, y \in \operatorname{dom} \psi$.
2. Hölder gradient: $\|\nabla f(x) - \nabla f(y)\| \leq L_\nu \|x - y\|^\nu$, $\nu \in [0, 1]$.
3. Unbiased stochastic oracle: $\mathbb{E}_\xi[g(x, \xi)] = \nabla f(x)$.
4. Bounded variance: $\mathbb{E}_\xi[\|g(x, \xi) - \nabla f(x)\|^2] \leq \sigma^2$.

Discussion:

- Most important example: ψ is $\{0, +\infty\}$ indicator of set Q.
- $\nu = 0$: better class than $\|\nabla f(x)\| \leq M$.
- Our methods require D and automatically adapt to σ, ν and L_ν.

Universal Stochastic Gradient Method

Input: Initial point $x_0 \in \operatorname{dom} \psi$, diameter D.

Set $H_0 = 0$ and compute $g_0 = g(x_0, \xi_0)$ for $\xi_0 \sim P_\xi$.

for $k \geq 0$ **do**

Compute $x_{k+1} = \operatorname{argmin}_{x \in \operatorname{dom} \psi} \{\langle g_k, x \rangle + \psi(x) + \frac{H_k}{2} \|x - x_k\|^2\}$.

Compute $g_{k+1} = g(x_{k+1}, \xi_{k+1})$ for $\xi_{k+1} \sim P_\xi$.

Update $H_{k+1} = H_k + [\hat{\beta}_{k+1} - \frac{H_k}{2} r_{k+1}^2]_+ / (D^2 + \frac{1}{2} r_{k+1}^2)$,

where $r_{k+1} = \|x_{k+1} - x_k\|$ and $\hat{\beta}_{k+1} = \langle g_{k+1} - g_k, x_{k+1} - x_k \rangle$.

- $\hat{\beta}_{k+1}$ is a stoch. estimate of symmetrized Bregman distance:

$$\hat{\beta}_f(x, y) = \langle \nabla f(y) - \nabla f(x), y - x \rangle = \beta_f(x, y) + \beta_f(y, x),$$

where $\beta_f(x, y) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$.

- Convergence rate for $\bar{x}_k = \frac{1}{k} \sum_{i=1}^{k} x_i$:

$$\mathbb{E}[F(\bar{x}_k)] - F^* \leq \frac{2\,\mathbb{E}[H_k] D^2}{k} \leq \inf_{\nu \in [0,1]} \frac{8 L_\nu D^{1+\nu}}{k^{(1+\nu)/2}} + \frac{4\sigma D}{\sqrt{k}}.$$

Formula for Step Size: Explanation in Deterministic Case

- Doing convergence analysis, we come to inequality

$$F(x_{k+1}) - F^* \leq \frac{H_k}{2}\rho_k^2 - \frac{H_{k+1}}{2}\rho_{k+1}^2 + U_k,$$
$$U_k = \frac{1}{2}(H_{k+1} - H_k)(\rho_{k+1}^2 + r_{k+1}^2) + \left[\beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2\right],$$

where $\rho_k = \|x_k - x^*\|$, $r_{k+1} = \|x_{k+1} - x_k\|$, $\beta_{k+1} = \beta_f(x_k, x_{k+1})$.

- A reasonable idea is to require that $H_k \leq H_{k+1}$, bound $\rho_{k+1} \leq D$ and $r_{k+1} \leq D$, and choose H_{k+1} to balance the two terms in U_k:

$$(H_{k+1} - H_k)D^2 = \left[\beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2\right]_+. \qquad (*)$$

- Solving (*), we get formula for H_{k+1} and simple recurrence

$$F(x_{k+1}) - F^* \leq \frac{H_k}{2}\rho_k^2 - \frac{H_{k+1}}{2}\rho_{k+1}^2 + 2(H_{k+1} - H_k)D^2.$$

Universal Stochastic Fast Gradient Method

Input: Initial point $x_0 \in \operatorname{dom}\psi$, diameter D.

Set $v_0 = x_0$, $H_0 = A_0 = 0$.

for $k \geq 0$ **do**

Set $a_{k+1} = k+1$, $A_{k+1} = A_k + a_{k+1}$, $\tau_k = a_{k+1}/A_{k+1}$.

Compute $y_k = (1-\tau_k)x_k + \tau_k v_k$ and $g_k^y = g(y_k, \xi_k^y)$ for $\xi_k^y \sim P_\xi$.

Set $v_{k+1} = \operatorname{argmin}_{x \in \operatorname{dom}\psi}\{a_{k+1}[\langle g_k^y, x\rangle + \psi(x)] + \frac{H_k}{2}\|x - v_k\|^2\}$.

$x_{k+1} = (1-\tau_k)x_k + \tau_k v_{k+1}$, $g_{k+1}^x = g(x_{k+1}, \xi_{k+1}^x)$, $\xi_{k+1}^x \sim P_\xi$.

Update $H_{k+1} = H_k + [A_{k+1}\hat{\beta}_{k+1} - \frac{H_k}{2} r_{k+1}^2]_+ / (D^2 + \frac{1}{2} r_{k+1}^2)$,

where $r_{k+1} = \|v_{k+1} - v_k\|$ and $\hat{\beta}_{k+1} = \langle g_{k+1}^x - g_k^y, x_{k+1} - y_k\rangle$

Convergence rate:

$$\mathbb{E}[F(x_k)] - F^* \leq \frac{4\,\mathbb{E}[H_k] D^2}{k(k+1)} \leq \inf_{\nu \in [0,1]} \frac{32 L_\nu D^{1+\nu}}{k^{(1+3\nu)/2}} + \frac{8\sigma D}{\sqrt{3k}}.$$

Choice of Diameter

- Our methods are not fully adaptive. They require knowledge of diameter D, as do all other existing adaptive stochastic methods.
- General adaptation to D is an interesting open question.
- However, for some important applications, D is known!

Example from Machine Learning

Regularized Empirical Risk Minimization:

$$\min_{x \in \mathbb{R}^n} \Big\{ \frac{1}{m} \sum_{i=1}^{m} f_i(x) + \frac{\lambda}{2} \|x\|^2 \Big\}, \qquad \text{(ERM)}$$

where f_i is the "loss function" for ith object (e.g., $f_i(x) = \ln(1 + e^{\langle a_i, x \rangle})$).

- We solve this problem for many values of λ and select the best model x using cross-validation.
- ERM is equivalent to

  $$\min_{x \in \mathbb{R}^n} \Big\{ \frac{1}{m} \sum_{i=1}^{m} f_i(x) : \|x\| \leq D/2 \Big\}.$$

  (Perfect problem for our methods!)
- Instead of searching for best λ, we can search for best D.

Open Questions

- Unbounded domain: $D \to R \geq \|x_0 - x^*\|$?
- Diagonal version (different step size for each coordinate)?
- Unconstrained nonconvex optimization: rates for $\|\nabla f(x_k)\|$?
- ...

Thank you!