ROYAL INSTITUTE
OF TECHNOLOGY

EP3260: Machine Learning Over Networks

Lecture 3: Centralized Convex ML

(part 2)

Hossein S. Ghadikolaei

Division of Network and Systems Engineering
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology, Stockholm, Sweden

https://sites.google.com/view/mlons/home

February 2020

Learning outcomes

- Recap of (deterministic) iterative algorithms for convex optimization
- Stochastic optimization
- Variance reduction techniques
- Convergence analysis

Outline

Recap of Lecture 2 and beyond!

Smooth problems (L-smooth, μ-strong convexity)

Gradient descent: minimize$_{\boldsymbol{w}\in\mathbb{R}^d}$ $f(\boldsymbol{w})$, $\mathcal{O}(1/k)$ for convex

Projected gradient descent: minimize$_{\boldsymbol{w}\in\mathcal{W}}$ $f(\boldsymbol{w})$, $\mathcal{O}(1/k)$ for convex

Steepest descent: minimize$_{\boldsymbol{w}\in\mathcal{W}}$ $f(\boldsymbol{w})$, large L/μ, $\mathcal{O}(1/k)$ for convex

Newton's methods: minimize$_{\boldsymbol{w}\in\mathcal{W}}$ $f(\boldsymbol{w})$, large L/μ

Acceleration methods: minimize$_{\boldsymbol{w}\in\mathcal{W}}$ $f(\boldsymbol{w})$, large L/μ, $\mathcal{O}(1/k^2)$ for convex

Nonsmooth problems

Subgradient methods: minimize$_{\boldsymbol{w}\in\mathbb{R}^d}$ $f(\boldsymbol{w})$, $\mathcal{O}(1/k)$ for convex

Proximal methods: minimize$_{\boldsymbol{w}\in\mathbb{R}^d}$ $g(\boldsymbol{w}) + h(\boldsymbol{w})$, $\mathcal{O}(1/k)$ for smooth f

Accelerated proximal methods: minimize$_{\boldsymbol{w}\in\mathbb{R}^d}$ $g(\boldsymbol{w}) + h(\boldsymbol{w})$, convex h, $\kappa = L/\mu$

update: $\boldsymbol{w}_{k+1} = \operatorname{prox}_{\alpha_k h}(\boldsymbol{v}_k - \alpha_k \nabla g(\boldsymbol{v}_k))$

momentum from prev. iteration: $\boldsymbol{v}_{k+1} = \boldsymbol{w}_{k+1} + \dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(\boldsymbol{w}_{k+1} - \boldsymbol{w}_k)$

Outline

Basic definitions

Convexity for differentiable function:

$$\nabla f(\boldsymbol{w}_1)^T(\boldsymbol{w}_2 - \boldsymbol{w}_1) \le f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1)$$

Strongly convexity:

$$f(\boldsymbol{w}_2) \ge f(\boldsymbol{w}_1) + \nabla f(\boldsymbol{w}_1)^T(\boldsymbol{w}_2 - \boldsymbol{w}_1) + \frac{\mu}{2}\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2^2$$

Smoothness:

$$f(\boldsymbol{w}_2) \le f(\boldsymbol{w}_1) + \nabla f(\boldsymbol{w}_1)^T(\boldsymbol{w}_2 - \boldsymbol{w}_1) + \frac{L}{2}\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2^2$$

Bounded error for initial guess: $\mathbb{E}\left[\|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|_2\right] \le R$

Lipschitz continuity (bounded gradients)

$$\|\boldsymbol{w}\|_2 \le D \Rightarrow \|\nabla f(\boldsymbol{w})\|_2 \le B$$

or $\|\boldsymbol{w}_1\|_2, \|\boldsymbol{w}_2\|_2 \le D \Rightarrow |f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1)| \le B\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$

Example

Consider Human Activity Recognition Using Smartphones dataset $\{(\boldsymbol{x}_i, y_i)\}_{i\in[N]}$

inputs: accelerometer and gyroscope sensors

output: moving (e.g., walking , running, dancing) or not (sitting or standing)

Consider logistic ridge regression: $\underset{\boldsymbol{w}}{\text{minimize}} \;\; f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2$
where $f_i(\boldsymbol{w}) = \log\left(1 + \exp\{-y_i \boldsymbol{w}^T \boldsymbol{x}_i\}\right)$

For classification, we can use the solution $\boldsymbol{w}^\star$ and compute $\mathrm{sign}({\boldsymbol{w}^\star}^T \boldsymbol{x})$

HW 2.1:

1) Is f Lipschitz continuous? If so, find a small B?

2) Is f_i smooth? If so, find a small L for f_i? What about f?

3) Is f strongly convex? If so, find a high μ?

Outline

Setting

- **Batch GD:** Let $f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w})$

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k) = \boldsymbol{w}_k - \frac{\alpha_k}{N}\sum_{i\in[N]} \nabla f_i(\boldsymbol{w}_k)$$

- **Stochastic gradient (SG) methods:**

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \; g(\boldsymbol{w}_k;\zeta_k) = \boldsymbol{w}_k - \alpha_k \widehat{\nabla} f(\boldsymbol{w}_k)$$

$\zeta_k \in [N]$, and $g(\boldsymbol{w}_k;\zeta_k)$ is a noisy version ("estimation") of $\nabla f(\boldsymbol{w}_k)$.

Method	**Per iteration cost**	**# iterations**
GD	Expensive (usually linear in N)	Usually few
SG	Very cheap, independent of N	Many

Main tradeoff: Per-iteration cost vs per-iteration improvement

Motivations for SG

Good theoretical guarantees: Consider strongly convex smooth f, then

- GD: $f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \mathcal{O}(\rho^k)$, $\rho \in (0, 1)$, so $N \log(1/\epsilon)$ total work for ϵ-optimality
- SG (basic version): $\mathbb{E}[f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star)] \leq \mathcal{O}(1/k)$, so $1/\epsilon$ total work for ϵ-optimality
- Compare $N \log(1/\epsilon)$ to $1/\epsilon$ for large N

Heavy computation

- Large scale optimization, $N \to \infty$, large matrix inversion

Heavy communication

- Bandwidth-limited distributed optimization

Security

- Revealing only a noisy gradient information

Nonconvex optimization and saddle points

Generic SG algorithm for decentralized optimization

A generic SG algorithm

Initialize $\boldsymbol{w}_1$
for $k = 1, 2, \ldots,$ **do**
  Generate a realization of the random variable ζ_k
  Compute a stochastic vector $g(\boldsymbol{w}_k; \zeta_k)$
  Choose step-size $\alpha_k > 0$
  Update $\boldsymbol{w}_{k+1} \leftarrow \boldsymbol{w}_k - \alpha_k \; g(\boldsymbol{w}_k; \zeta_k)$
end for

- Problem: $\underset{\boldsymbol{w}}{\text{minimize}} \; f(\boldsymbol{w}) = \frac{1}{N} \sum_{i \in [N]} f_i(\boldsymbol{w})$.

  decentralized vs distributed

- Examples of stochastic vector

  Gradient for one sample: $\nabla f_{\zeta_k}(\boldsymbol{w}_k)$

  Gradient for a mini-batch: $\frac{1}{N_k} \sum_{i \in [N_k]} \nabla f_{\zeta_k, i}(\boldsymbol{w}_k)$

  Preconditioned mini-batch gradient: $H_k \frac{1}{N_k} \sum_{i \in [N_k]} \nabla f_{\zeta_k, i}(\boldsymbol{w}_k)$

Outline

Smoothness

Observe that $\boldsymbol{w}_{k+1}$ depends only on ζ_k, and assume i.i.d. $(\zeta_k)_k$

$\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})]$: expectation of $f(\boldsymbol{w}_{k+1})$ wrt the distribution of ζ_k only

f being L-smooth implies that the generic SG algorithm satisfies for all $k \in \mathbb{N}$

$$\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})] - f(\boldsymbol{w}_k) \leq$$
$$- \underbrace{\alpha_k \nabla f(\boldsymbol{w}_k)^T \mathbb{E}_{\zeta_k}\left[g(\boldsymbol{w}_k;\zeta_k)\right]}_{\text{expected decrease}} + \underbrace{\frac{1}{2}\alpha_k^2 L \mathbb{E}_{\zeta_k}\left[\|g(\boldsymbol{w}_k;\zeta_k)\|_2^2\right]}_{\text{noise}}$$

If $g(\boldsymbol{w}_k;\zeta_k)$ is an unbiased estimate of $\nabla f(\boldsymbol{w}_k)$, then

$$\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})] - f(\boldsymbol{w}_k) \leq -\alpha_k \|\nabla f(\boldsymbol{w}_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\zeta_k}\left[\|g(\boldsymbol{w}_k;\zeta_k)\|_2^2\right] \quad (1)$$

Some useful assumptions

- The sequence $\{\boldsymbol{w}_k\}$ is contained in an open set over which f is bounded below by a scalar f_{inf}

- There exist scalars $c_0 \geq c > 0$ s.t. for all $k \in \mathbb{N}$

$$\nabla f(\boldsymbol{w}_k)^T \mathbb{E}_{\zeta_k}\left[g(\boldsymbol{w}_k; \zeta_k)\right] \geq c\|\nabla f(\boldsymbol{w}_k)\|_2^2 \tag{2a}$$

$$\|\mathbb{E}_{\zeta_k}\left[g(\boldsymbol{w}_k; \zeta_k)\right]\|_2 \leq c_0\|\nabla f(\boldsymbol{w}_k)\|_2 \tag{2b}$$

- There exist scalars $M \geq 0$ and $M_V \geq 0$ s.t. for all $k \in \mathbb{N}$

$$\mathrm{Var}_{\zeta_k}\left[g(\boldsymbol{w}_k; \zeta_k)\right] \leq M + M_V\|\nabla f(\boldsymbol{w}_k)\|_2^2 \tag{3}$$

For unbiased gradient estimator: $c = c_0 = 1$

(2) and (3) imply (HW 2.2: find α and β.)

$$\mathbb{E}_{\zeta_k}\left[\|g(\boldsymbol{w}_k; \zeta_k)\|_2^2\right] \leq \alpha + \beta\|\nabla f(\boldsymbol{w}_k)\|_2^2$$

An important tradeoff

Generic SG algorithm on L-smooth function satisfies

$$\begin{aligned}
&\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})] - f(\boldsymbol{w}_k) \\
&\qquad \le -c\alpha_k \|\nabla f(\boldsymbol{w}_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\zeta_k}\left[\|g(\boldsymbol{w}_k;\zeta_k)\|_2^2\right] \\
&\qquad \le -\left(c - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \|\nabla f(\boldsymbol{w}_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M
\end{aligned} \tag{4}$$

Proof: see the board

Convergence of SG depends on the balance between blue and red terms

Outline

Strongly convex f and fixed step-size

Theorem 1

For all $k \in \mathbb{N}$ and constant step-size $\alpha_k = \alpha$ satisfying

$$0 < \alpha \leq \frac{c}{LM_G}\,, \tag{5}$$

the expected optimality gap satisfies

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] &\leq \frac{\alpha LM}{2\mu c} + (1-\alpha\mu c)^{k-1}\left(f(\boldsymbol{w}_1) - f^\star - \frac{\alpha LM}{2\mu c}\right) \\ &\xrightarrow{k\to\infty} \frac{\alpha LM}{2\mu c}\end{aligned} \tag{6}$$

where $M_G = M_V + c_0^2$.

If $g(\boldsymbol{w}_k; \zeta_k)$ is unbiased estimate of $\nabla f(\boldsymbol{w}_k)$, then $c = 1$, we may assume $M_G = 1$ and retrieve $\alpha \in (0, 1/L]$ of GD

Proof

Additional notes

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] - \frac{\alpha L M}{2\mu c} \leq (1-\alpha\mu c)^{k-1}\left(f(\boldsymbol{w}_1) - f^\star - \frac{\alpha L M}{2\mu c}\right)$$

Fast convergence to a neighborhood of the optimal value, but noise in the gradient prevented further progress (convergence to an ambiguity ball)

Optimality gap $\frac{\alpha L M}{2\mu c}$

Contraction constant after k iteration $(1-\alpha\mu c)^{k-1}$

A simple modification: run SG with a fixed step-size, and after convergence halve the step-size and run SG again, ...

- How $E[f(\boldsymbol{w}_k)]$ against k behaves now?
- No sub-optimality gap
- Each time the step-size is cut in half, double the number of iterations are required
- **Effective convergence rate $\mathcal{O}(1/k)$**, why?

Strongly convex f and diminishing step-size

Theorem 2

For all $k \in \mathbb{N}$ and diminishing step-size α_k satisfying

$$\alpha_k = \frac{\beta}{\gamma + k}, \text{ for some } \beta > \frac{1}{\mu c} \text{ and } \gamma > 0 \text{ s.t. } \alpha_1 \leq \frac{c}{L M_G},$$

the expected optimality gap satisfies

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] \leq \frac{\nu}{\gamma + k} \tag{7}$$

where

$$\nu := \max\left\{\frac{\beta^2 L M}{2\left(\beta\mu c - 1\right)}, (\gamma + 1)\left(f(\boldsymbol{w}_1) - f^\star\right)\right\}$$

Usually first term of ν determines the asymptotic convergence of $\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right]$

Proof

Additional notes

New step-size parameter $\beta > \frac{1}{\mu c}$:

Sensitive to overestimation of μ

higher $\mu \rightarrow$ smaller $\beta \rightarrow$ slower convergence rate

Constant step-size and mini-batch vs diminishing step-size

For mini-batch, define $g(\boldsymbol{w}_k; \zeta_k) = \frac{1}{N_m} \sum_{i \in [N_m]} \nabla f_{\zeta_k, i}(\boldsymbol{w}_k)$

Mini-batch with small constant $\alpha > 0$,

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] \leq \frac{\alpha L M}{2\mu c N_m} + (1 - \alpha\mu c)^{k-1} \left(f(\boldsymbol{w}_1) - f^\star - \frac{\alpha L M}{2\mu c N_m}\right)$$

Simple SG with small constant α / N_m, (cheap iterations, many iterations)

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] \leq \frac{\alpha L M}{2\mu c N_m} + \left(1 - \frac{\alpha\mu c}{N_m}\right)^{k-1} \left(f(\boldsymbol{w}_1) - f^\star - \frac{\alpha L M}{2\mu c N_m}\right)$$

Convex f and diminishing step-size

- Notations:
  - $\mathbb{E}[g(\boldsymbol{w};\zeta_k)|\boldsymbol{w}_k] \in \partial f(\boldsymbol{w}_k)$: noisy unbiased sub-gradient of convex f
  - $f_{\text{best}}(\boldsymbol{w}_k) = \min\left(f(\boldsymbol{w}_1), \ldots, f(\boldsymbol{w}_k)\right)$
  - $\mathbb{E}\left[\|g(\boldsymbol{w}_k;\zeta_k)\|_2^2\right] \leq G^2$ for all k, and $\sup_{\boldsymbol{w}\in\mathcal{W}} \mathbb{E}\left[\|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|_2^2\right] \leq R^2$

Theorem 3

Under some mild conditions and for square summable but not summable step-size, we have convergence in expectation

$$\mathbb{E}\left[f_{\text{best}}(\boldsymbol{w}_k) - f^\star\right] \leq \frac{R^2 + G^2 \sum_{i\in[k]} \alpha_i^2}{2\sum_{i\in[k]} \alpha_i}$$

and for any arbitrary $\epsilon, \delta > 0$, we have convergence in probability:

$$\Pr\left(f_{\text{best}}(\boldsymbol{w}_k) - f^\star \geq \epsilon\right) \leq \delta$$

Proof

Convex f and diminishing step-size

Theorem 4

For convex L-smooth function f, i.i.d. stochastic gradient of variance bound σ^2, and diminishing step-size $\alpha_k = \frac{1}{L+\gamma^{-1}}$, where $\gamma = \frac{R}{G}\sqrt{\frac{2}{k}}$, we have

$$\mathbb{E}\left[f\left(\frac{1}{k} \sum_{i \in [k]} \boldsymbol{w}_k \right) - f^\star \right] \leq R\sqrt{\frac{2\sigma^2}{k}} + \frac{LR^2}{k} \tag{8}$$

Proof: see [Bubeck 2015, Theorem 6.3]

Improved gain for mini-batch of size N_m: $\sigma^2 \to \sigma^2 / N_m$

Non-convex objective function

Theorem 5

With fixed step-size as of (5), for all $K \in \mathbb{N}$, we have

$$\mathbb{E}\left[\sum_{k\in[K]} \|\nabla f(\boldsymbol{w}_k)\|_2^2\right] \leq \frac{K\alpha LM}{c} + \frac{2(f(\boldsymbol{w}_1) - f_{\inf})}{c\alpha} \tag{9}$$

and therefore

$$\mathbb{E}\left[\frac{1}{K}\sum_{k\in[K]} \|\nabla f(\boldsymbol{w}_k)\|_2^2\right] \leq \frac{\alpha LM}{c} + \frac{2(f(\boldsymbol{w}_1) - f_{\inf})}{Kc\alpha} \xrightarrow{K\to\infty} \frac{\alpha LM}{c} \tag{10}$$

Proof: Recursively $\forall k \in [K]$, take total expectation from (4), use (5), observe

$$f_{\inf} - f(\boldsymbol{w}_1) \leq \mathbb{E}\left[f(\boldsymbol{w}_{K+1})\right] - f(\boldsymbol{w}_1) \leq -\tfrac{1}{2}c\alpha \textstyle\sum_{k\in[K]} \mathbb{E}\left[\|\nabla f(\boldsymbol{w}_k)\|_2^2\right] + \tfrac{1}{2}K\alpha^2 LM.$$

$f_{\inf}$ is not necessarily $f^\star$

SG spends increasingly more time in regions where the objective function has a “relatively” small gradient. Also usual tradeoff on step-size.

Non-convex objective function

Theorem 6

With square summable but not summable step-size, we have for any $K \in \mathbb{N}$

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \|\nabla f(\boldsymbol{w}_k)\|_2^2\right] < \infty \tag{11}$$

and therefore

$$\mathbb{E}\left[\frac{1}{\sum_{k\in[K]} a_k} \sum_{k\in[K]} \alpha_k \|\nabla f(\boldsymbol{w}_k)\|_2^2\right] \xrightarrow{K\to\infty} 0 \tag{12}$$

Proof: **HW 2.3**!

The expected gradient norm cannot stay bounded away from zero

Foods for thought

1. Recall from Theorem 1 that SG with a constant step-size converges linearly to an ambiguity ball whose radius is determined by the variance of the gradient noise

   Observe that taking $N_k > 1$ samples "with replacement" implies multiplying the radius of the ambiguity ball by N_k^{-1} and conclude "doubling the batch size cuts the error in half"

   Observe that taking $N_k > 1$ samples "without replacement" implies multiplying the radius of the ambiguity ball by $\frac{N-N_k}{NN_k}$

   Modify the generic SG algorithm with a dynamic batch size.

   Can we recover the linear convergence rate to $\boldsymbol{w}^\star$? Linear in terms of iterations or workload (effect computations)? Note the increasing cost of iterations (due to larger N_k with k)

2. Often in practice, features (inputs, $\boldsymbol{x} \in \mathcal{X}$) of dimension d are very sparse (at most $z \ll d$ non-zero elements)

   Modify SG method to have $\mathcal{O}(z)$ cost per iteration instead of the original $\mathcal{O}(d)$

   Can we do that for all objective functions? What about an SVM classifier?

3. In decentralized/distributed computing, we may have a high communication overhead to exchange $\boldsymbol{w}_k$ among workers. Can we use the vanilla SG method to tradeoff the costs between computation and communication?

Outline

Basic Ideas

- Increasing mini-batch size
- Reducing step-size α_k to reduce the radius of the ambiguity ball
- Reducing variance of sample of X by using a sample from random variable Y with a known expectation

$$Z_a = a(X - Y) + \mathbb{E}[Y]$$

$\mathbb{E}[Z_a] = a\mathbb{E}[X] + (1 - a)\mathbb{E}[Y]$

$\mathrm{var}(Z_a) = a^2 \left(\mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X, Y)\right)$

$a = 1$: no bias, $a < 1$: reduced variance at the expense of potential bias

Stochastic variance reduced gradient (SVRG)

SVRG (Johnson&Zhang, 2013; Zhang et. al., 2013)

Inputs: Epoch length T, number of epochs K
for $k = 1, 2, \ldots, K$ **do**
  Compute all gradients and store $\widetilde{\nabla} f := \frac{1}{N} \sum_{i \in [N]} \nabla f_i(\widetilde{\boldsymbol{w}}_k)$
  Initialize $\boldsymbol{w}_{k,0} \leftarrow \widetilde{\boldsymbol{w}}_k$
  for t=1,...,T **do**
    Sample ζ_k uniformly from $[N]$
    $\boldsymbol{w}_{k,t} \leftarrow \boldsymbol{w}_{k,t-1} - \alpha_k \left(\nabla f_{\zeta_k}(\boldsymbol{w}_{k,t-1}) - \nabla f_{\zeta_k}(\widetilde{\boldsymbol{w}}_k) + \widetilde{\nabla} f \right)$
  end for
  Update $\widetilde{\boldsymbol{w}}_{k+1} \leftarrow \boldsymbol{w}_{k,T}$
end for
Return: $\widetilde{\boldsymbol{w}}_{K+1}$

- $X = \boldsymbol{w}$, $Y = \widetilde{\boldsymbol{w}}$, with a known average (blue step)
- One memory, two gradients per inner loop
- Linear convergence rate (given a sufficiently large T)

Proof

Stochastic average gradient (SAG)

SAG (Schmidt&Le Roux&Bach, 2012, 2017)

for $k = 1, 2, \ldots,$ **do**

Sample ζ_k uniformly from $[N]$ and observe $\nabla f_i(\boldsymbol{w}_k)$

Update for all $i \in [N]$, $\hat{g}_i(\boldsymbol{w}_k) = \begin{cases} \nabla f_i(\boldsymbol{w}_k), & \text{if } i = \zeta_k \\ \hat{g}_i(\boldsymbol{w}_{k-1}), & \text{otherwise} \end{cases}$

Update $\boldsymbol{w}_{k+1} \leftarrow \boldsymbol{w}_k - \frac{\alpha_k}{N} \sum_{i\in[N]} \hat{g}_i(\boldsymbol{w}_k)$

end for

- Almost same convergence rate (and same proof) as of SVRG
- A memory of size N
- Biased gradient estimates: $\mathbb{E}\left[\frac{1}{N}\sum_{i\in[N]} \hat{g}_i(\boldsymbol{w}_k)\right] = \frac{1}{N}\sum_{i\in[N]} \nabla f_i(\boldsymbol{w}_k)$ does not hold necessarily
- Table averaging representation and SAGA extension

Which algorithm to choose?

CA1: Closed-form solution vs iterative approaches

Consider $\boldsymbol{x}^\star = \underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\ \frac{1}{N}\sum_{i\in[N]} \|\boldsymbol{w}^T\boldsymbol{x}_i - \boldsymbol{y}_i\|^2 + \lambda\|\boldsymbol{w}\|_2^2$ for dataset $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$

1) Find a closed-form solution for this problem

2) Consider "Communities and Crime" dataset ($N = 1994$, $d = 128$) and find the optimal linear regressor from the closed-form expression

3) Repeat 2) for "Individual household electric power consumption" dataset ($N = 2075259$, $d = 9$) and observe the scalability issue of the closed-form expression

4) How would you address even bigger datasets?

CA2: Deterministic/stochastic algorithms in practice

Consider logistic ridge regression $f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2$ where $f_i(\boldsymbol{w}) = \log\left(1 + \exp\{-y_i\boldsymbol{w}^T\boldsymbol{x}_i\}\right)$ for "Individual household electric power consumption" dataset

1) Solve the optimization problem using GD, stochastic GD, SVRG, and SAG

2) Tune a bit hyper-parameters (including λ)

3) Compare these solvers in terms complexity of hyper-parameter tunning, convergence time, convergence rate (in terms of # outer-loop iterations), and memory requirement

Some references

- L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," SIAM Review, 2018.

- S. Bubeck, "Convex optimization: Algorithms and complexity," Foundations and Trends in Machine Learning, 2015.

- S. Boyd and A. Mutapcic, "Stochastic subgradient methods," Lecture Notes for EE364b, Stanford University, 2018.

- R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," NIPS, 2013.

- L. Zhang, M. Mahdavi, and R. Jin, "Linear convergence with condition number independent access of full gradients," NIPS 2013.

- M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," Mathematical Programming, 2017.

Outline

Proof sketch for Theorem 1

Use (4), Polyak-Lojasiewicz inequality (a consequence of strong convexity) and (5), and observe that

$$\begin{aligned}
\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})] - f(\boldsymbol{w}_k) &\le -\left(c - \frac{1}{2}\alpha L M_G\right)\alpha\|\nabla f(\boldsymbol{w}_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\
&\le -\frac{1}{2}\alpha c\|\nabla f(\boldsymbol{w}_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\
&\le -\alpha\mu c\,(f(\boldsymbol{w}_k) - f^\star) + \frac{1}{2}\alpha^2 LM
\end{aligned}$$

Subtract $f^\star$ from both sides, take total expectation, and rearrange:

$$\mathbb{E}\left[f(\boldsymbol{w}_{k+1}) - f^\star\right] \le (1-\alpha\mu c)\,\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] + \frac{1}{2}\alpha^2 LM$$

Make it a contraction inequality (as $0 < \alpha\mu c \le \frac{\mu c^2}{LM_G} \le \frac{\mu}{L} \le 1$)

$$\mathbb{E}\left[f(\boldsymbol{w}_{k+1}) - f^\star\right] - \frac{\alpha LM}{2\mu c} \le (1-\alpha\mu c)\left(\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] - \frac{\alpha LM}{2\mu c}\right).$$

Return

Proof sketch for Theorem 2

First observe that $\alpha_k L M_G \leq \alpha_1 L M_G \leq c$. Use (4) and Polyak-Lojasiewicz inequality and show that

$$\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})] - f(\boldsymbol{w}_k) \leq -\alpha_k \mu c \left(f(\boldsymbol{w}_k) - f^\star\right) + \frac{1}{2}\alpha_k^2 L M$$

Subtract $f^\star$ from both sides, take total expectation, and rearrange:

$$\mathbb{E}\left[f(\boldsymbol{w}_{k+1}) - f^\star\right] \leq (1 - \alpha_k \mu c)\, \mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] + \frac{1}{2}\alpha_k^2 L M$$

Now prove by induction and use inequality $k^2 \geq (k+1)(k-1)$

Return

Proof sketch for Theorem 3

Use convexity of f $\left(f^\star - f(\boldsymbol{w}_k) \geq \mathbb{E}[g(\boldsymbol{w};\zeta_k)|\boldsymbol{w}_k]^T(\boldsymbol{w}^\star - \boldsymbol{w}_k)\right)$ to show

$$\mathbb{E}\left[\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|_2^2 \mid \boldsymbol{w}_k\right] \leq \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|_2^2 - 2\alpha_k\left(f(\boldsymbol{w}_k) - f^\star\right) + \alpha_k^2 G^2$$

Take expectation nd apply recursively to show

$$\mathbb{E}\left[\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|_2^2\right] \leq \mathbb{E}[\|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|_2^2] - 2\sum_{i\in[k]}\alpha_i\left(\mathbb{E}[f(\boldsymbol{w}_i)] - f^\star\right) + G^2\sum_{i\in[k]}\alpha_i^2$$

Conclude that for square summable but not summable step-size, $\min_{i\in[k]}\mathbb{E}[f(\boldsymbol{w}_i)] \to f^\star$

Use Jensen's inequality and concavity of minimum to show convergence in expectation $\mathbb{E}[f_{\mathsf{best}}(\boldsymbol{w}_k)] = \mathbb{E}[\min_{i\in[k]} f(\boldsymbol{w}_i)] \leq \min_{i\in[k]}\mathbb{E}[f(\boldsymbol{w}_i)] \to f^\star$

Use Markov's inequality to show convergence in probability:
$\Pr(f_{\mathsf{best}}(\boldsymbol{w}_k) - f^\star \geq \epsilon) \leq \frac{\mathbb{E}[f_{\mathsf{best}}(\boldsymbol{w}_k) - f^\star]}{\epsilon}$

Return

Linear convergence of SVRG

Variance decomposition:

$$\mathbb{E}\left[\|\boldsymbol{w} - \mathbb{E}\left[\boldsymbol{w}\right]\|_2^2\right] \leq \mathbb{E}\left[\|\boldsymbol{w}\|_2^2\right] - \|\mathbb{E}\left[\boldsymbol{w}\right]\|_2^2 \leq \mathbb{E}\left[\|\boldsymbol{w}\|_2^2\right]$$

Show

$$\mathbb{E}_{\zeta_k}\left[\left\|\nabla f_{\zeta_k}(\boldsymbol{w}_{k,t-1}) - \nabla f_{\zeta_k}(\widetilde{\boldsymbol{w}}_k) + \widetilde{\nabla} f\right\|_2^2\right] \leq 4L\left(f(\boldsymbol{w}_{k,t-1}) + f(\widetilde{\boldsymbol{w}}_k) - 2f^\star\right)$$

Use the inner-loop iteration and bound $\mathbb{E}_{\zeta_k}\left[\|\boldsymbol{w}_{k,t} - \boldsymbol{w}^\star\|_2^2\right]$. You may need to use convexity of f

Sum $\mathbb{E}_{\zeta_k}\left[\|\boldsymbol{w}_{k,t} - \boldsymbol{w}^\star\|_2^2\right]$ over the inner loop ($t \in [T]$) and cancel some terms from both sides

Show and use for every outer iteration to observe the linear convergence rate: if $a < ba + c$ for $b \in (0,1)$, then

$$a - \frac{c}{1-b} \leq b\left(a - \frac{c}{1-b}\right)$$

Return