Optimization methods
Lecture 8: Introduction to stochastic gradient methods

Alexandr Katrutsa

Modern State of Artificial Intelligence Masters Program
Moscow Institute of Physics and Technology

Brief reminder of the previous lecture

- Conjugate gradient method
- Heavy-ball method
- Accelerated gradient method

What do we know?

- Deterministic first-order methods

What do we know?

- Deterministic first-order methods
- Fast methods for different classes of problems

What do we know?

- Deterministic first-order methods
- Fast methods for different classes of problems

Questions

- How the methods will change if the randomness will be introduced in problems?
- How to measure convergence in that case?

Why do we need randomness?

- If the number of variables is huge, the explicit computing of the gradient can be hard

Why do we need randomness?

- If the number of variables is huge, the explicit computing of the gradient can be hard
- Stochastic gradient estimate can be sufficient for solving problem at the appropriate level

Why do we need randomness?

- If the number of variables is huge, the explicit computing of the gradient can be hard
- Stochastic gradient estimate can be sufficient for solving problem at the appropriate level
- Sometimes given parameters of the problem are inexact

How the randomness can be introduced?

- The known data in the problem is random variables with known distributions

$$\begin{aligned} &\min x_1 + x_2 \\ \text{s.t. } & w_1 x_1 + x_2 \geq 0 \\ & w_2 x_1 + x_2 \geq 0 \\ & x_{1,2} \geq 0, \end{aligned}$$

  where $w_1 \sim \mathcal{U}[0,4]$, $w_2 \sim \mathcal{U}[2,3]$

How the randomness can be introduced?

- The known data in the problem is random variables with known distributions

$$\begin{aligned} &\min x_1 + x_2 \\ \text{s.t. } &w_1 x_1 + x_2 \geq 0 \\ &w_2 x_1 + x_2 \geq 0 \\ &x_{1,2} \geq 0, \end{aligned}$$

  where $w_1 \sim \mathcal{U}[0, 4]$, $w_2 \sim \mathcal{U}[2, 3]$
- Objective function is an expected value of some other function

$$\min f(\mathbf{x}) := \mathbb{E}_\omega[F(\mathbf{x}, \omega)]$$

How the randomness can be introduced?

- The known data in the problem is random variables with known distributions

$$\begin{aligned} & \min x_1 + x_2 \\ \text{s.t. } & w_1 x_1 + x_2 \geq 0 \\ & w_2 x_1 + x_2 \geq 0 \\ & x_{1,2} \geq 0, \end{aligned}$$

  where $w_1 \sim \mathcal{U}[0,4]$, $w_2 \sim \mathcal{U}[2,3]$
- Objective function is an expected value of some other function

$$\min f(\mathbf{x}) := \mathbb{E}_\omega[F(\mathbf{x}, \omega)]$$

  - A particular case

$$\min \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x})$$

SAA vs SA

- Stochastic approximation (SA)

SAA vs SA

- Stochastic approximation (SA)
  - Generate ω^k i.i.d.

SAA vs SA

- Stochastic approximation (SA)
  - Generate ω^k i.i.d.
  - Compute stochastic gradient $G(\mathbf{x}, \omega^k)$

SAA vs SA

- Stochastic approximation (SA)
  - Generate ω^k i.i.d.
  - Compute stochastic gradient $G(\mathbf{x}, \omega^k)$
  - Use it in stochastic gradient descent

SAA vs SA

- Stochastic approximation (SA)
  - Generate ω^k i.i.d.
  - Compute stochastic gradient $G(\mathbf{x}, \omega^k)$
  - Use it in stochastic gradient descent
- Sample average approximation (SAA)

SAA vs SA

- Stochastic approximation (SA)
  - Generate ω^k i.i.d.
  - Compute stochastic gradient $G(\mathbf{x}, \omega^k)$
  - Use it in stochastic gradient descent
- Sample average approximation (SAA)
  - Generate N samples $\omega_1, \ldots, \omega_N$

SAA vs SA

- Stochastic approximation (SA)
  - Generate ω^k i.i.d.
  - Compute stochastic gradient $G(\mathbf{x}, \omega^k)$
  - Use it in stochastic gradient descent
- Sample average approximation (SAA)
  - Generate N samples $\omega_1, \ldots, \omega_N$
  - Compute estimate of the objective $\hat{f}_N(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} F(\mathbf{x}, \omega_i)$

SAA vs SA

- Stochastic approximation (SA)
  - Generate ω^k i.i.d.
  - Compute stochastic gradient $G(\mathbf{x}, \omega^k)$
  - Use it in stochastic gradient descent
- Sample average approximation (SAA)
  - Generate N samples $\omega_1, \ldots, \omega_N$
  - Compute estimate of the objective $\hat{f}_N(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} F(\mathbf{x}, \omega_i)$
  - Minimize $\hat{f}_N$ instead of the original function f

Problem statement

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x})$$

- $f_i(\mathbf{x})$ may be nonconvex
- n may be of the order 10^6 and higher
- N is also may be huge

Example 1

- Hutchinson trace estimator

$$\mathrm{trace}(\mathbf{A}) = \mathrm{trace}(\mathbf{A}\mathbf{I}) = \mathrm{trace}(\mathbf{A}\mathbb{E}_{\mathbf{z}}\mathbf{z}\mathbf{z}^\top) = \mathbb{E}_{\mathbf{z}}(\mathbf{z}^\top\mathbf{A}\mathbf{z}),$$

  where $\mathbf{z}$ is a vector from standard normal distribution or from the Rademacher distribution
- Expected value is replaced with the unbiased estimate $\hat{f}_N$ similar to SAA approach
- Minimize $\hat{f}_N$ for fixed $\mathbf{z}_i$

Example 2

- Classification problem
- Loss function ℓ is additive by the samples of the training set

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{w}|\mathbf{x}_i)$$

- Interpretation as the empirical risk minimization or ground truth distribution approximation

Stochastic gradient descent (SGD)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{h}_k,$$

where

- $\mathbf{h}_k = f'_{i_k}(x_k)$, $i_k \in \{1, \dots, N\}$ is selected randomly
- $\mathbf{h}_k = \frac{1}{|\mathcal{I}_k|} \sum\limits_{i \in \mathcal{I}_k} f'_i(\mathbf{x}_k)$, $\mathcal{I}_k \subset \{1, \dots, N\}$ is some subset of indices usually of fixed size $|\mathcal{I}_k| = m$

Properties

1. Unbiased gradient estimate

$$\mathbb{E}[\mathbf{h}_k] = f'(\mathbf{x}_k)$$

2. Large variance

Convergence

Theorem

Let f be convex, L-smooth function. Then if SGD generates directions $\mathbf{h}_k$ such that $\mathrm{Var}(\mathbf{h}_k) \leq \sigma^2$ and $\alpha_k \leq \frac{1}{L}$ then

$$\mathbb{E}[f(\bar{\mathbf{x}}_k)] - f^* \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{\alpha_k k} + \frac{\alpha_k \sigma^2}{2}.$$

In particular, after $k = \frac{(\sigma^2 + L\|\mathbf{x}^* - \mathbf{x}_0\|_2^2)^2}{\varepsilon^2}$ iterations if $\alpha_k = \frac{1}{\sqrt{k}}$ we get the solution with accuracy 2ε.

General approach to variance reduction

- Assume X_ω gives unbiased estimate of the parameter x: $\mathbb{E}_\omega[X_\omega] = x$

General approach to variance reduction

- Assume X_ω gives unbiased estimate of the parameter x: $\mathbb{E}_\omega[X_\omega] = x$
- Assume $Z_\omega = X_\omega - Y_\omega$ such that $\mathbb{E}_\omega[Y_\omega] \approx 0$

General approach to variance reduction

- Assume X_ω gives unbiased estimate of the parameter x: $\mathbb{E}_\omega[X_\omega] = x$
- Assume $Z_\omega = X_\omega - Y_\omega$ such that $\mathbb{E}_\omega[Y_\omega] \approx 0$
- Then $\mathbb{E}_\omega[X_\omega] = \mathbb{E}_\omega[Z_\omega] = x$

General approach to variance reduction

- Assume X_ω gives unbiased estimate of the parameter x: $\mathbb{E}_\omega[X_\omega] = x$
- Assume $Z_\omega = X_\omega - Y_\omega$ such that $\mathbb{E}_\omega[Y_\omega] \approx 0$
- Then $\mathbb{E}_\omega[X_\omega] = \mathbb{E}_\omega[Z_\omega] = x$
- $\mathrm{Var}(Z_\omega) = \mathrm{Var}(X_\omega) + \mathrm{Var}(Y_\omega) - 2\mathrm{Cov}(X_\omega, Y_\omega) \ll \mathrm{Var}(X_\omega)$ if Y_ω highly correlates with X_ω

General approach to variance reduction

- Assume X_ω gives unbiased estimate of the parameter x: $\mathbb{E}_\omega[X_\omega] = x$
- Assume $Z_\omega = X_\omega - Y_\omega$ such that $\mathbb{E}_\omega[Y_\omega] \approx 0$
- Then $\mathbb{E}_\omega[X_\omega] = \mathbb{E}_\omega[Z_\omega] = x$
- $\mathrm{Var}(Z_\omega) = \mathrm{Var}(X_\omega) + \mathrm{Var}(Y_\omega) - 2\mathrm{Cov}(X_\omega, Y_\omega) \ll \mathrm{Var}(X_\omega)$ if Y_ω highly correlates with X_ω

The recipe to reduce the variance

Find the estimate Y, such that

1. Its expected value is close to 0
2. It highly correlates with given estimate X

Stochastic average gradient (Schmidt, Le Roux, Bach 2013)

- Initialization x_0 and $g_i^0 = x_0, i = \{1, \ldots, N\}$
- In the k-th iteration, one selects some index i_k and updates $g_{i_k}^k = f'_{i_k}(x_k)$
- $x_{k+1} = x_k - \alpha_k \frac{1}{N} \sum\limits_{i=1}^{N} g_i^k$
- More convenient notation

$$x_{k+1} = x_k - \alpha_k \left(\frac{1}{N} g_{i_k}^{(k+1)} - \frac{1}{N} g_{i_k}^k + \frac{1}{N} \sum_{i=1}^{N} g_i^k \right)$$

Variance reduction

- $X = g_{i_k}^{(k+1)}$ and $\mathbb{E}_\omega[X] = f'(x_k)$
- $Y = g_{i_k}^k - \sum_{i=1}^{N} g_i^k$ and $\mathbb{E}_\omega[Y] \neq 0$
- $\|X - Y\|_2 = \|(g_{i_k}^{(k+1)} - g_{i_k}^k) + \sum_{i=1}^{N} g_i^k\|_2 \to 0,\ k \to \infty$
- Variance of the result estimate goes to 0

Convergence for convex and L-smooth function

Theorem

Let f_i be differentiable and L-smooth, $\bar{x}^{(k)} = \frac{1}{k}\sum_{i=0}^{k-1} x_i$, $\alpha_k = \frac{1}{16L}$ and initialization

$$g_i^0 = f_i'(x_0) - f'(x_0),\ i = 1, \dots, N$$

gives

$$\mathbb{E}[f(\bar{x}^{(k)})] - f(x^*) \leq \frac{48n}{k}(f(x_0) - f^*) + \frac{128L}{k}\|x_0 - x^*\|_2^2$$

Comparison

- SAG
$$\frac{48n}{k}(f(x_0) - f^*) + \frac{128L}{k}\|x_0 - x^*\|_2^2$$
The first item depends on n!
- GD
$$\frac{L\|x_0 - x^*\|_2^2}{k}$$
- SGD
$$\frac{\|x_0 - x^*\|_2^2 + \sigma^2}{2\sqrt{k}}$$

Convergence for L-smooth and μ-strongly convex function

Theorem

If there are the same assumptions that were used in the theorem about convex L-smooth function, then the foloowing estimate holds

$$\mathbb{E}[f(\bar{x}^{(k)})] - f(x^*) \leq \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8n}\right\}\right)^k \Big(\frac{3}{2}(f(x_0) - f^*) + \frac{4L}{n}\|x^* - x_0\|_2^2\Big)$$

- Adapt to the strong convexity
- Analogue of the SGD
- SGD gives only $\mathcal{O}(1/\sqrt{k})$ convergence rate

Remarks

- SAG requires careful tuning of settings
- Initial approximation is better to derive from one epoch of SGD and storing g_i^0
- Choice of α_k

SAGA (Defazio, Bach, Lacoste-Julien 2014)

The analogue of SAG, but

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \left(\mathbf{g}_{i_k}^{(k+1)} - \mathbf{g}_{i_k}^{k} + \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i^k \right)$$

- Unbiased estimate: $\mathbb{E}[Y] = 0$
- Variance is higher than in SAG
- The analysis of variance reduction is the same
- Can be generalized in the composite problems
- Convergence estimates are the same as they are for SAG
- Implementation details are the same as for SAG

SVRG (Johnson, Zhang 2013)

- Initialization $\bar{x}_0$
- For $k = 1, 2, \dots$
  - $\bar{x} = \bar{x}_0$
  - $\bar{\mu} = f'(\bar{x})$
  - $x_0 = \bar{x}_0$
  - For $m = 1, \dots, l$
    - Random choice of $i_m \in \{1, \dots, N\}$
    - $$x_{m+1} = x_m - \alpha(f'_{i_m}(x_m) - f'_{i_m}(\bar{x}) + \bar{\mu})$$
  - $\bar{x}_0 = x_l$

Features of SVRG

- Analogue of SAGA
- The proof is simpler
- It depends on the number of epoch

Drawbacks of variance reduction methods

- They require exact gradient computations
- They depend on other parameters
- No universal way to run them

Summary

- Stochastic estimate of the gradient helps in many cases

Summary

- Stochastic estimate of the gradient helps in many cases
- SGD and its properties

Summary

- Stochastic estimate of the gradient helps in many cases
- SGD and its properties
- Variance reduction methods