Advance Stochastic Gradient with Variance Reduction

Jingchang Liu

December 7, 2017

University of Science and Technology of China

Table of Contents

Introductions

Formulations

Optimization problems

$$\min f(w), \qquad f(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Stochastic gradient descent
At each iteration $t = 1, 2, \cdots$, draw i_t randomly from $\{1, \cdots, n\}$

$$w^t = w^t - \eta_t \nabla f_{i_t}\left(w^t\right)$$

Unified formulation
ζ is a random variable.

$$w^{t+1} = w^t - \eta_t g(w^t, \zeta_t)$$

Estimation

Stochastic gradient

$$\nabla f_{i_t}(w^t) \to \frac{1}{n}\sum_{i=1}^{n} f_i(w^t)$$

Unbiased

$$\mathbb{E}\left\{\nabla f_{i_t}(w^t)\right\} = \frac{1}{n}\sum_{i=1}^{n} f_i(w^t)$$

Variance Reduce(VR)

control variates	antithetic variates
important sampling	stratified sampling

Control Variates

Control variates

Introduction
Unknown parameter μ, assume we have a static $X : \mathbb{E}X = \mu$, another r.v. Y, such that $\mathbb{E}Y = \tau$ is a known value, define a new r.v.

$$\bar{X} = X + c(Y - \tau)$$

Properties

- Unbias: $\mathbb{E}\bar{X} = \mathbb{E}X = \mu$
- Variance: $Var(\bar{X}) = Var(X) + c^2 Var(Y) + 2cCov(X, Y)$
  Optimal coefficient: $c^* = -\frac{Cov(X,Y)}{Var(Y)}$
- Simply:
  - $\bar{X} = X - Y + \tau$,if $cov(X, Y) > 0$
  - $\bar{X} = X + Y - \tau$,if $cov(X, Y) < 0$

Control variates for stochastic gradient

VR gradient

- Former: $v_k = \nabla f_{i_k}(w_{k-1})$
- Case 1: $v_k = \nabla f_{i_k}(w_{k-1}) - \nabla h_{i_k}(w_{k-1}) + \mathbb{E}\nabla h_{i_k}(w_{k-1})$
- Case 2: $v_k = \nabla f_{i_k}(w_{k-1}) - \nabla f_{i_k}(\tilde{w}) + \tilde{v}$

Methods

- SAGA: $\nabla f_{i_k}(\tilde{w})$ is stored in the table.
- SVRG: $\nabla f_{i_k}(\tilde{w})$ is calculated after a specific number of iterations.
- $\lim_{k \to 0} \mathbb{E} \|v_k\|^2 = 0$
- SAGA. SVRG will convergence under fixed stepsize.

Antithetic Sampling

antithetic variates

Two r.v. X_i, X_j id, $\mathbb{E}X_i = \mu, \mathbb{E}X_j = \mu$.
As $\mathbb{E}\left\{\frac{1}{2}(X_i + X_j)\right\} = \mu$ use $\frac{1}{2}(X_i + X_j)$ to estimate μ

Formulations

- if X and Y are independent,

$$\begin{aligned} Var(\frac{1}{2}(X_i + X_j)) &= \frac{1}{4}Var(X_i + X_j) = \frac{1}{4}\left\{Var(X_i) + Var(X_j)\right\} \\ &= \frac{1}{4} \times 2Var(X_i) = \frac{1}{2}Var(X_i) \end{aligned}$$

- if X and Y are negative correlation,

$$Var(\frac{1}{2}(X_i + X_j)) = \frac{1}{4}\{Var(X_i) + Var(X_j) + 2Cov(X_i, X_j)\} \leq \frac{1}{2}Var(X_i)$$

- if $X_j = 2\mu - X_i$, then $Var(\frac{1}{2}(X_i + X_j)) = Var(\mu) = 0$

antithetic variates for stochastic gradient

logistic regression

$$\nabla f_i(w) = \frac{e^{-y_i \cdot x_i^{'} w}}{1 + e^{-y_i \cdot x_i^{'} w}} y_i x_i^{'}$$

Formulations

$$\mathbb{E} \left\| \nabla f_i(w) + \nabla f_j(w) \right\|^2 = \mathbb{E} \left\| \nabla f_i(w) \right\|^2 + \mathbb{E} \left\| \nabla f_j(w) \right\|^2 + 2\mathbb{E} \left\langle \nabla f_i(w), \nabla f_j(w) \right\rangle$$

$$\begin{aligned} \mathbb{E} \left\langle \nabla f_i(w), \nabla f_j(w) \right\rangle &= \mathbb{E} \left\langle \frac{e^{-y_i \cdot x_i^{'} w}}{1 + e^{-y_i \cdot x_i^{'} w}} y_i x_i^{'}, \frac{e^{-y_j \cdot x_j^{'} w}}{1 + e^{-y_j \cdot x_j^{'} w}} y_j x_j^{'} \right\rangle \\ &\geq -\mathbb{E} \left\| \frac{e^{-y_i \cdot x_i^{'} w}}{1 + e^{-y_i \cdot x_i^{'} w}} y_i x_i^{'} \right\| \left\| \frac{e^{-y_j \cdot x_j^{'} w}}{1 + e^{-y_j \cdot x_j^{'} w}} y_j x_j^{'} \right\| \end{aligned}$$

if and only if $y_i x_i^{'} \parallel y_j x_j^{'}$, equal hold.

SDCA

Derivation

$f\left(w\right)=\frac{1}{n}\sum\limits_{i=1}^{n}f_i\left(w\right)+\frac{\lambda}{2}\left\|w\right\|^2$ equals to

$$\begin{array}{rcl} P\left(y,z\right) & = & \frac{1}{n}\sum\limits_{i=1}^{n}f_i\left(z_i\right)+\frac{\lambda}{2}\left\|y\right\|^2 \\ s.t. & & y=z_i, i=1,2,\cdots,n \end{array}$$

$L\left(y,z,\alpha\right)=P\left(y,z\right)+\frac{1}{n}\sum\limits_{i=1}^{n}\alpha_i\left(y-z_i\right)$

$$\begin{array}{rcl} D\left(\alpha\right) & = & \inf\limits_{y,z}L\left(y,z,\alpha\right) \\ & = & \frac{1}{n}\sum\limits_{i=1}^{n}\inf\limits_{z_i}\left\{f_i\left(z_i\right)-\alpha_i z_i\right\}+\inf\limits_{y}\left\{\frac{\lambda}{2}\left\|y\right\|^2+\frac{1}{n}\sum\limits_{i=1}^{n}\alpha_i y\right\} \\ & = & \frac{1}{n}\sum\limits_{i=1}^{n}-f_i^*\left(-\alpha_i\right)-\frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum\limits_{i=1}^{n}\alpha_i\right\|^2 \end{array}$$

SDCA

Formulation and relationships

$$\min f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) + 0.5\lambda w' w$$

$$\alpha_i^* = -\frac{1}{\lambda n}\nabla f_i(w^*) \qquad w^t = \sum_{i=1}^{n} \alpha_i^t$$

Update

$$\alpha_l^t = \begin{cases} \alpha_l^{t-1} - \eta_t(\nabla f_i(w^{t-1}) + \lambda n \alpha_l^{t-1}) & l = i \\ \alpha_l^{t-1} & l \neq i \end{cases}$$

$$\begin{aligned} w^t &= w^{t-1} + \left(\alpha_i^t - \alpha_i^{t-1}\right) \\ &= w^{t-1} - \eta_t(\nabla f_i(w^{t-1}) + \lambda n \alpha_l^{t-1}) \end{aligned}$$

$\lambda n \alpha_l^{t-1}$ is antithetic to $\nabla f_i(w^{t-1})$, $\nabla f_i(w^{t-1}) + \lambda n \alpha_l^{t-1} \to 0$ as $t \to \inf$

Stratified Sampling

Stratified sampling

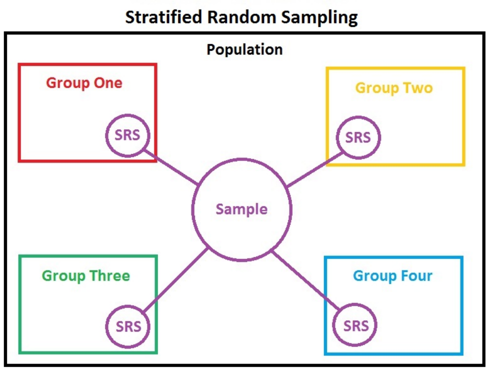

Figure 1: Stratified sampling

Group 1	Group 2	Group 3	Group 4
$\nabla f_{11}, \cdots, \nabla f_{1n_1}$	$\nabla f_{21}, \cdots, \nabla f_{2n_2}$	$\cdots$	$\nabla f_{L1}, \cdots, \nabla f_{Ln_L}$

Stratified sampling

Principles

- homogenous within-groups.
- heterogenous between the groups.

Stratified sample size

- Proportional: $\frac{b_h}{b} = \frac{n_h}{n} = W_h$
- Neyman: $b_h = b\frac{W_h S_h}{\sum\limits_{h=1}^{L} W_h S_h} = b\frac{N_h S_h}{\sum\limits_{h=1}^{L} N_h S_h}$

Apply to stochastic gradient

- for the same labels y, cluster x, to stratify.
- $(x_i, y_i) \to \nabla f_i(w; x_i, y_i)$

Important Sampling

Important Sampling

- Uniform sampling: $\nabla f(w^t) = \sum_{i=1}^{n} \boxed{\frac{1}{n}} \nabla f_i(w^t)$
- Important sampling: $\nabla f(w^t) = \sum_{i=1}^{n} \frac{\nabla f_i(w)}{np_i^t} \boxed{p_i^t}$, $\sum_{i=1}^{n} p_i^t = 1 \quad t = 1, 2, \cdots$

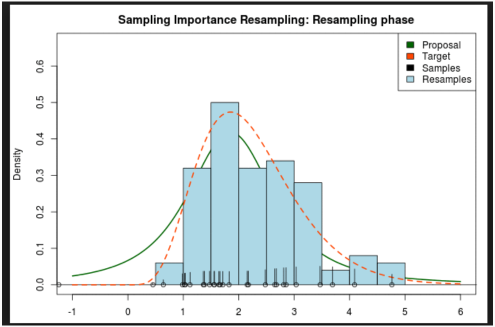

Figure 2: Important sampling

Important Sampling for Stochastic Gradient

$$\min_{p^t} E\left\|\frac{\nabla f_{i_t}(w^t)}{np^t_{i_t}}\right\|^2 = \min_{p^t} \frac{1}{n^2}\sum_{i=1}^{n}\frac{\left\|\nabla f_i(w^t)\right\|^2}{p_i^t} \geq \frac{1}{n^2}\left(\sum_{i=1}^{n}\left\|\nabla f_i(w^t)\right\|\right)^2$$

$$p_i^t = \frac{\|\nabla f_i(w^t)\|}{\sum_{j=1}^{n}\|\nabla f_j(w^t)\|}$$

if $f_i(w)$ is L_i-Lipschitz, then $\|\nabla f_i(w)\| \leq L_i$,

$$p_i^t = \frac{L_i}{\sum_{j=1}^{n} L_j}$$

Experiments

Stratified Sampling

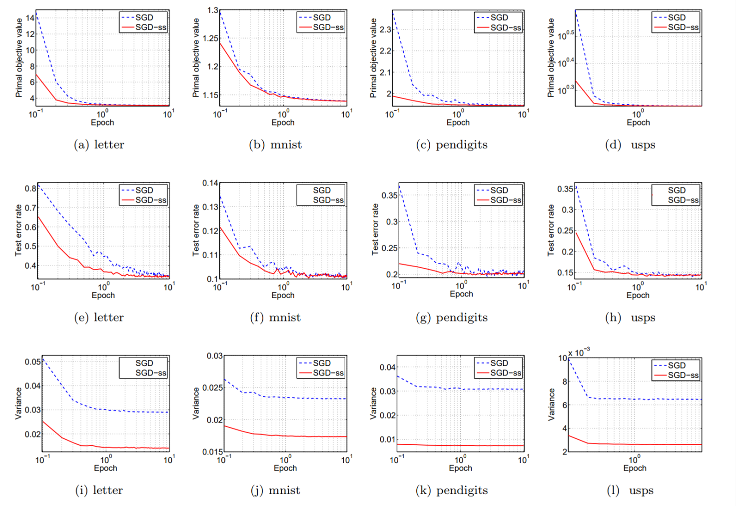

Figure 3: multi-class logistic regression (convex) on letter, mnist, pendigits, and usps.

Important sampling

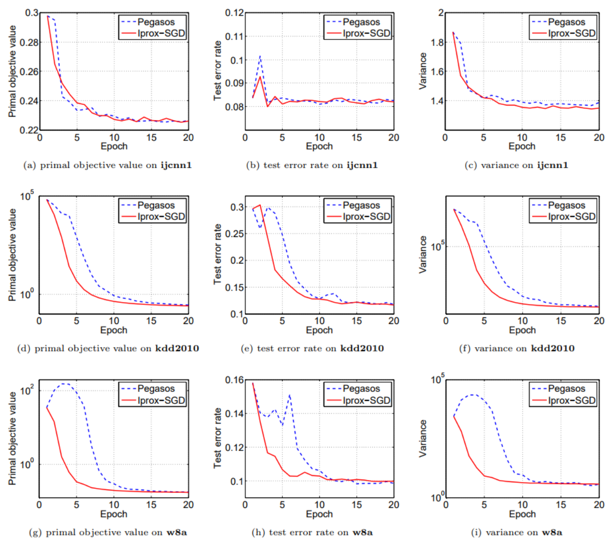

Figure 4: SVM on several datasets

Conclusions

conclusions

- VR base on optimize variables, such as SDCA. SVRG, can make the variance convergence to 0.
- VR base on samples, can significantly reduce the variance.
- Constructing related variates is crucial.
- Different VR methods can be combined, but how to need our efforts.

Q & A