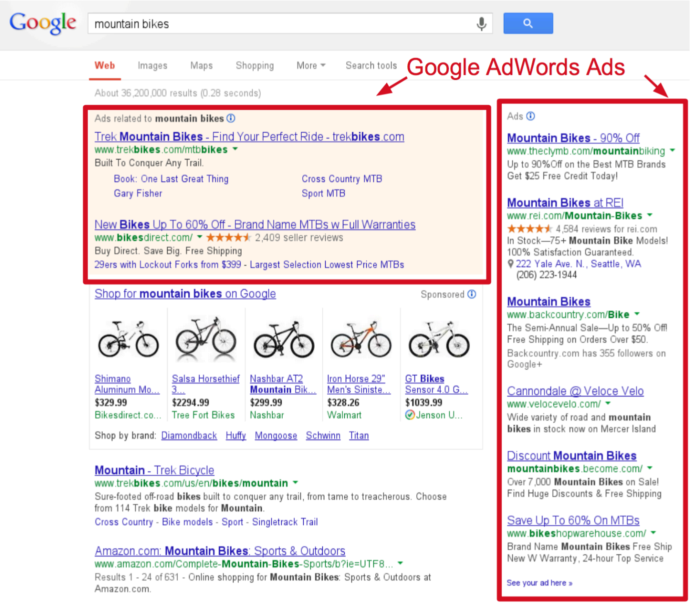
Review: The Statistics of Streaming Sparse Regression

Shih-Kang Chao
Department of Statistics
Purdue University

June, 2016


Google
mountain bikes
Web Images Maps Shopping More Search tools
Google AdWords Ads
About 36,200,000 results (0.28 seconds)
Ads related to mountain bikes
Trek Mountain Bikes - Find Your Perfect Ride - trekbikes.com
www.trekbikes.com/mtbbikes
Built To Conquer Any Trail.
Book: One Last Great Thing
Cross Country MTB
Gary Fisher
Sport MTB
New Bikes Up To 60% Off - Brand Name MTBs w Full Warranties
www.bikesdirect.com/ 2,409 seller reviews
Buy Direct. Save Big. Free Shipping
29ers with Lockout Forks from $399 - Largest Selection Lowest Price MTBs
Shop for mountain bikes on Google
Sponsored
Shimano Aluminum Mo... $329.99 Bikesdirect.co...
Salsa Horsethief 3... $2294.99 Tree Fort Bikes
Nashbar AT2 Mountain Bik... $299.99 Nashbar
Iron Horse 29" Men's Siniste... $328.26 Walmart
GT Bikes Sensor 4.0 G... $1039.99 Jenson U...
Shop by brand: Diamondback Huffy Mongoose Schwinn Titan
Mountain - Trek Bicycle
www.trekbikes.com/us/en/bikes/mountain
Sure-footed off-road bikes built to conquer any trail, from tame to treacherous. Choose from 114 Trek bike models for Mountain.
Cross Country - Bike models - Sport - Singletrack Trail
Amazon.com: Mountain Bikes: Sports & Outdoors
www.amazon.com/Complete-Mountain-Bikes-Sports/b?ie=UTF8...
Results 1 - 24 of 631 - Online shopping for Mountain Bikes: Sports & Outdoors at Amazon.com.
Ads
Mountain Bikes - 90% Off
www.theclymb.com/mountainbiking
Up to 90%Off on the Best MTB Brands
Get $25 Free Credit Today!
Mountain Bikes at REI
www.rei.com/Mountain-Bikes
4,584 reviews for rei.com
In Stock—75+ Mountain Bike Models!
100% Satisfaction Guaranteed.
222 Yale Ave. N., Seattle, WA
(206) 223-1944
Mountain Bikes
www.backcountry.com/Bike
The Semi-Annual Sale—Up to 50% Off!
Free Shipping on Orders Over $50.
Backcountry.com has 355 followers on Google+
Cannondale @ Veloce Velo
www.velocevelo.com/
Wide variety of road and mountain bikes in stock now on Mercer Island
Discount Mountain Bikes
mountainbikes.become.com/
Over 7,000 Mountain Bikes on Sale!
Find Huge Discounts & Free Shipping
Save Up To 60% On MTBs
www.bikeshopwarehouse.com/
Brand Name Mountain Bikes Free Ship
New W Warranty, 24-hour Top Service
See your ad here »

Ad Click Prediction

- Ads shown in response to a query
- The search engine only gets *paid* if the user clicks the ad. Only relevant ads are to be shown
- Predicting the probability of a click for a specific ad in response to a specific query is the key
- Mathematical framework: $y_t = \mathbf{1}\{\text{Click an ad}\}$, $P(y_t|q_t, x_t)$. q_t : user's query, x_t : features of an ad.

Challenges

- High-dimension: billions of unique features (or model coefficients)
- Immediate response: billions of predictions per day serving live traffic
- Big data: billions of training examples

Other applications:

- Astrophysics
- Environmental sensor networks
- Console logs mining in large-scale datacenter

General Streaming Data Setting

Assume a generalized linear model $\mathbb{E}[y_t] = g(x_t^\top w^*)$, where $g : \mathbb{R} \to \mathbb{R}$ is a known link function

In each step:

1. Observe covariates $x_t \in \mathbb{R}^d$
2. Make a prediction $\hat{y}_t \in \mathbb{R}$ (using preobtained coefficients $w_t \in \mathbb{R}^d$)
3. Observe realization $y_t \in \mathbb{R}$
4. Update w_t to w_{t+1}

Ad click example: $g(u) = \frac{1}{1+e^{-u}}$. Logistic regression.

Stochastic Gradient Decent

$f_t(w)$:loss functions; e.g. $f_t(w) = \frac{1}{2}(y_t - w^\top x_t)^2$.
$w^* := \arg\min_{w\in\mathbb{R}^d} \mathbb{E}[f_t(w)]$

Stochastic Gradient Decent (SGD):

Asymptotics

$$w_{t+1} = w_t - \frac{1}{\eta t}\nabla f_{t+1}(w_t), \tag{1.1}$$

where $\eta > 0$ is some step size

- $\nabla f_t(w_t)$ is random (depending on y_t, x_t, w_t)
- SGD is not robust to the choice of η, an alternative method is proposed in TRA14

ℓ_1 penalized SGD

- SGD does not encourage sparsity
- SGD is equivalent to the following adaptive mirror decent update for $t = 1, 2...$

$$
\begin{aligned}
\theta_t &= \sum_{s=1}^{t-1} \nabla f_{s+1}(w_s) \\
w_t &= \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{\eta}{2} \sum_{s=1}^{t-1} \|w_s - w\|_2^2 + w^\top \theta_t \right\} \qquad (2.1)
\end{aligned}
$$

ℓ_1 penalized SGD

Introducing ℓ_1 norm into (2.1)

$$\theta_t = \sum_{s=1}^{t-1} \nabla f_{s+1}(w_s)$$

$$w_t = \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{\eta}{2} \sum_{s=1}^{t-1} \|w_s - w\|_2^2 + w^\top \theta_t + \lambda\sqrt{t+1}\|w\|_1 \right\}.$$

This is called Streaming Sparse Regression (SSR).

Algorithm 1 Streaming sparse regression. S_λ denotes the soft-thresholding operator: $S_\lambda(x) = 0$ if $|x| < \lambda$, and $x - \lambda\,\mathrm{sign}(x)$ otherwise.

Input: sequence of loss functions $f_1, \ldots, f_T$
Output: parameter estimate w_T
Algorithm parameters: η, λ, ϵ

$\theta_1 = 0$
for $t = 1$ **to** T **do**
  $\lambda_t \leftarrow \lambda\sqrt{t+1}$
  $w_t \leftarrow \frac{1}{\epsilon + \eta\,(t-1)} S_{\lambda_t}(\theta_t)$ ▷ sparsification step
  $\theta_{t+1} = \theta_t - [\nabla f_t(w_t) - \eta\, w_t]$ ▷ gradient step
end for
return w_T

- Accommodate big data ($T \to \infty$)
- Fast computation: proximity operator S_λ has a closed form

Theoretical Properties

Measure of performance: w_t : Algorithm 1, $\hat{w}_t$: Algorithm 2

$$\text{Regret}(w^*) := \sum_{t=1}^{T} (f_t(w_t) - f_t(w^*)),$$

Parameter error: $\|\hat{w}_T - w^*\|_2^2$.

Bound for Regret

Theorem (1)

Under regularity assumptions, $\|\nabla f_t(w)\|_\infty \leq B$ *for some* $B > 0$, $\alpha > 0$ *is a restricted strictly convexity constant. If* Assumptions

$$\lambda = \frac{3B}{2}\sqrt{\log\left(\frac{6d\log_2(2T)}{\delta}\right)},$$

$\eta = \alpha/2$, $\epsilon = 0$ *are applied in Algorithm 1. Then for any* $\delta > 0$, *with probability* $1-\delta$, *we have*

$$\text{Regret}(w^*) = O\left(\frac{kB^2}{\alpha}\log\left(\frac{d\log(T)}{\delta}\right)\log(T)\right). \tag{2.2}$$

$\alpha\uparrow$: greater curvature; $k = |\text{supp}(w^*)|$

Parameter Error...

Algorithm 2 Streaming sparse regression with averaging. S_λ denotes the soft-thresholding operator: $S_\lambda(x) = 0$ if $|x| < \lambda$, and $x - \lambda\,\mathrm{sign}(x)$ otherwise.

Input: sequence of functions $f_1, \ldots, f_T$
Output: parameter estimate $\hat{w}_T$
Algorithm parameters: η, λ, ϵ

$\hat{w}_0 = 0$, $\theta_1 = 0$
for $t = 1$ **to** T **do**
  $\lambda_t \leftarrow t^{\frac{3}{2}}\lambda$
  $w_t \leftarrow \frac{1}{\epsilon + \eta\, t(t-1)/2} S_{\lambda_t}(\theta_t)$ $\triangleright$ sparsification step
  $\theta_{t+1} \leftarrow \theta_t - t\left[\nabla f_t(w_t) - \eta\, w_t\right]$ $\triangleright$ gradient step
  $\hat{w}_t \leftarrow \left(1 - \frac{2}{t+1}\right)\hat{w}_{t-1} + \frac{2}{t+1} w_t$ $\triangleright$ averaging step
end for
return $\hat{w}_T$

- Better handle correlated noise features
- Generate sharper parameter error

Parameter Error

Theorem (2)

Under the same conditions as Theorem 1, but now using Algorithm 2 and set

$$\lambda = \frac{3B}{2}\sqrt{\log\left(\frac{6d\log_2(2T^3)}{\delta}\right)}.$$

Then for any $\delta > 0$, with probability $1-\delta$, we have $\operatorname{supp}(\hat{w}_T) \subset S$ *and*

$$\|\hat{w}_T - w^*\|_2^2 = O\left(\frac{kB^2}{\alpha^2 T}\log\left(\frac{d\log(T)}{\delta}\right)\right). \tag{2.3}$$

- Achieve the batch lasso minimax rate

Idea of Proof: Regret

A general bound (Orabona et al., 2015):

$$\sum_{t=1}^{T}(w_t - u)^\top \nabla f_t(w_t)$$
$$\leq \psi_T(u) + \sum_{t=1}^{T} D_{\psi_t^*}(\theta_{t+1}\|\theta_t) + \sum_{t=1}^{T}[\psi_{t-1}(w_t) - \psi_t(w_t)],$$

for the general problem

$$w_t = \arg\min_{w\in\mathbb{R}^d}\{\psi_t(w) + w^\top\theta_t\}, \quad \theta_t = \sum_{s=1}^{t-1}\nabla f_{s+1}(w_s), \tag{2.4}$$

where $\psi_t^*(u) = \sup_v \left(v^\top u - \psi_t(v)\right)$, and the Bregman divergence

$$D_{\psi_t^*}(\theta_{t+1}\|\theta_t) = \psi_t^*(\theta_{t+1}) - \psi_t^*(\theta_t) - \langle\nabla\psi_t^*(\theta_t), \theta_{t+1} - \theta_t\rangle$$

Idea of Proof: Regret

- If losses f_t are convex, then the regret can be bounded

$$f_t(w_t) - f_t(u) \le (w_t - u)^\top \nabla f_t(w_t)$$

- For the current setting of ψ

$$\sum_{t=1}^{T} (w_t - w^*)^\top \nabla f_t(w_t) \le \Omega_0 + \Lambda + Q,$$

$$\Omega_0 = \frac{\epsilon}{2}\|w^*\|_2^2 + \frac{1}{2}\sum_{t=1}^{T} \frac{\|\nabla f_t(w_t)\|_2^2}{\epsilon + \eta t} = O_P(k \log T),$$

$$\Lambda = \sum_{t=1}^{T} \lambda(\sqrt{t-1} - \sqrt{t})(\|w_t\|_1 - \|w^*\|_1) = \lambda k \log T,$$

$$Q = \frac{\eta}{2}\sum_{t=1}^{T} \|w_t - w^*\|_2^2 \text{ (combined with strictly convexity)}$$

Idea of Proof: Parameter Error

Online-to-batch conversion: If Regret(w^*)= $O_P(Q(T))$ for some function $Q(T)$, then

$$\mathcal{L}(\hat{w}_T) - \mathcal{L}(w^*) = O_P\left(\frac{Q(T)}{T}\right), \text{ with } \hat{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t.$$

▸ Definition of $\mathcal{L}$

Then by strong convexity, for any w,

$$\frac{1}{2}\|w - w^*\|_2^2 \leq \frac{1}{\alpha}(\mathcal{L}(w) - \mathcal{L}(w^*)).$$

Irrepresentatble Noise Features

Assumption

For any $\tau \in \mathbb{R}^d$ with $supp(\tau) \subset S$ and any $j \notin S$,

$$\left|\mathrm{Cov}(x_t^j, \tau \cdot x_t)\right| \leq \rho \frac{\alpha}{\sqrt{k}} \|\tau\|_2$$

for some constant $0 \leq \rho < 1/\sqrt{24}$, where α is the strong convexity parameter of the expected loss, and $|S| = k$.

The irrepresentability allows for a little nonzero covariance, which is weaker than the orthogonality condition Assumptions

Parameter Error with Irrepresentability

Theorem (3)

Under regularity conditions but replacing orthogonality condition (assumption 4) with the irrepresentability condition. Then, using Algorithm 2, for any $\delta > 0$, for an appropriate setting of λ we have

$$\|\hat{w}_T - w^*\|_2^2 = O\left(\frac{1}{1 - 24\rho^2}\frac{kB^2}{\alpha^2 T}\log\left(\frac{d\log T}{\delta}\right)\right)$$

with probability $1 - \delta$.

If $\rho = 0$, go back to Theorem 2

Simulation Setting

$f_t(w) = h(y_t - x_t^\top w)$, where

$$h(y) = \begin{cases} y^2/2 & : |y| \leq C \\ C \cdot (|y| - C/2) & : |y| \geq C. \end{cases}$$

which is more robust to outliers than L_2 loss

1. Linear regression: $y_t = x_t^\top w^* + v_t$, $v_t \sim \mathcal{N}(0, \sigma^2)$, $x_t \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.8^{|i-j|}$
2. Logistic regression: x_t is a random sign vector and $y_t \in \{0, 1\}$. $P(y_t = 1|x_t) = \frac{1}{1+\exp(-x_t^\top w^*)}$

$d = 100000$. The first 100 entries of w^* were nonzero.

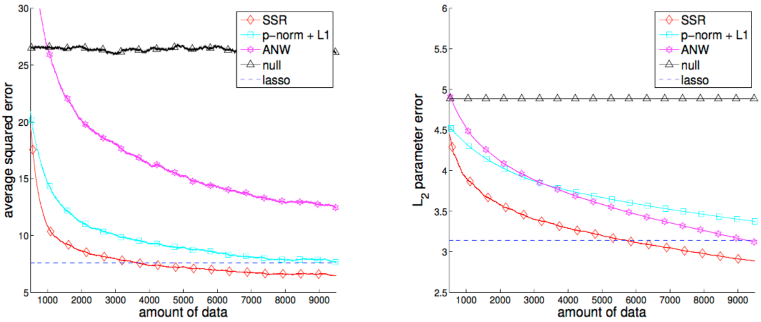
Simulation Setting

Competing models:

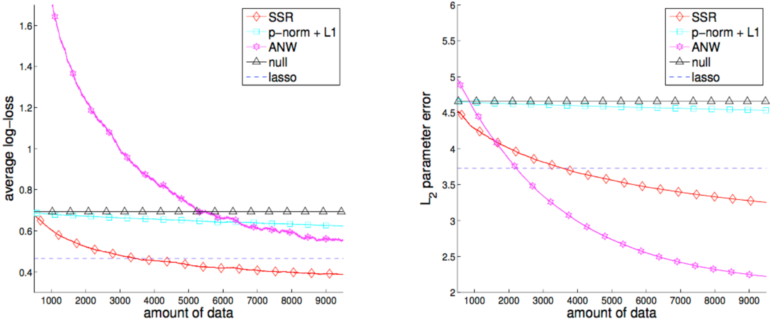
- p-norm regularized dual averaging (p-norm+L_1)(Shalev-Shwartz and Tewari, 2011)
- Epoch-based algorithm of Agarwal et al. (2012): optimal asymptotic rates
- Oracle: Using all data at once. Applying Lasso with `glmnet` (can handle at most 2500 data points before it crashes in MATLAB)

SGD performs very bad and is not included

Linear Reg., Correlated Features



Logistic Regression



Asymptotics of SGD

$b_t := \mathbb{E}[w_t] - w^*$; $V_t := Var\,(w_t)$.
$\mathcal{I} := -\mathbb{E}[\nabla_w^2 f_t(w^*)]$, σ is a dispersion measure of y_t,
$\sum_{t=1}^{\infty} t^{-1} c_t < \infty$

SGD

Theorem (4)

Under regularity conditions (Assumption 4.1 in TRA14), the asymptotic bias of the SGD satisfies

$$b_t = (\mathbf{I} - (\eta t)^{-1}\sigma\mathcal{I}(w^*))b_{t-1} + c_t \tag{4.1}$$

Moreover, if there exists $\alpha > 0$ such that $2\psi\mathcal{I} - \mathbf{I}/\alpha$ is positive definite, then the asymptotic variance

$$(\eta t)V_t \to \alpha\sigma^2(2\alpha\psi\mathcal{I} - \mathbf{I})^{-1}\mathcal{I}. \tag{4.2}$$

Asymptotic normality follows by Sacks (1958)

Recursive Expression for SSR

(Not in the literature)
By Karush-Kuhn-Tucker condition, we have

$$w_t = \underbrace{w_{t-1} - \frac{1}{\eta(t-1)}\nabla f_t(w_{t-1})}_{\text{SGD}} - \underbrace{\frac{\lambda}{\eta(t-1)}(\sqrt{t+1}\hat{\kappa}_t + \sqrt{t}\hat{\kappa}_{t-1})}_{\text{from } \ell_1 \text{ norm subgradient}}$$

$\|\hat{\kappa}_t\|_\infty \leq 1$, $(\hat{\kappa}_t)_i = \text{sign}((w_t)_i)$ if $(w_t)_i \neq 0$.

- $\lambda = 0$, back to classical SGD
- Bias is accumulative
- Step size η affects the asymptotic distribution

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Stochasticoptimization and sparse statistical recovery: Optimal algorithms for high dimensions. *Advances in Neural Information Processing Systems*, pages 1538–1546.

Orabona, F., Crammer, K., and Cesa-Bianchi, N. (2015). A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99:411–435.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Annals of Mathematical Statistics*, 29(2):373–405.

Shalev-Shwartz, S. and Tewari, A. (2011). Stochastic methods for l1-regularized loss minimization. *The Journal of Machine Learning Research*, 12:1865–1892.

1. **Statistical Sparsity:** There is a fixed expected loss function $\mathcal{L}$ such that
$$\mathbb{E}\left[f_t \mid f_1, \ldots, f_{t-1}\right] = \mathcal{L} \quad \text{for} \quad t = 1, 2, \ldots$$
Moreover, the minimizer w^* of the loss $\mathcal{L}$ satisfies $\|w^*\|_1 \leq R$ and $\mathrm{supp}(w^*) = S$, where $|S| \leq k$. Define the set of candidate weight vectors:
$$\mathcal{H} \stackrel{\text{def}}{=} \{w : \|w\|_1 \leq R, \mathrm{supp}(w) \subseteq S\}.$$
We note that $\mathcal{H}$ is not directly available to the statistician, because she does not know S.
2. **Strong Convexity in Expectation:** There is a constant $\alpha > 0$ such that $\mathcal{L}(w) - \frac{\alpha}{2}\|w_{[S]}\|_2^2$ is convex. Recall that, for an arbitrary vector w, $w_{[S]}$ denotes the coordinates indexed by S and $w_{[\neg S]}$ denotes the remaining coordinates.
3. **Bounded Gradients:** The gradients ∇f_t satisfy $\|\nabla f_t(w)\|_\infty \leq B$ for all $w \in \mathcal{H}$.
4. **Orthogonal Noise Features:** For our simplest results, we assume that the noise gradients are mean-zero for all $w \in \mathcal{H}$: more precisely, for all $i \notin S$ and all $w \in \mathcal{H}$, we have $\nabla \mathcal{L}(w)_i = 0$. In Section 2.3 below, we discuss how we can relax this condition into an irrepresentability condition.

Theorem 1 Idea of Proof