Consistently better than SGD

Best regret bound

Ready to use

---

1. The Least Squares problem

Objective function:
\[ \min_w \sum_{i=1}^{n} \frac{1}{2} ||y_i - w^T x_i||^2_2, \]
where \( x_i \in \mathbb{R}^d, y_i \in \mathbb{R}^1 \) and \( w \in \mathbb{R}^d \).

Linear regression

2. When Least Squares problem meets Large Scale

Use closed form solution: \( w = (X^T X)^{-1} X^T y \)?

It’s \( O(nd^2) \)!

Note: Any algorithm with time complexity greater than \( O(nd) \) is not applicable in large scale high dimension cases.

Stochastic Gradient Descent (SGD) with \( O(d) \) time each iteration is an appealing approach, which takes the form:

\[ w_{t+1} = \arg \min_w \frac{1}{t} \sum_{i=1}^t \frac{1}{2} ||y_i - w^T x_i||^2_2, \]

where \( l(w, x_i, y_i) = \frac{1}{2} ||y_i - w^T x_i||^2_2 \) is the loss function at step \( t \), denoted as \( l(w) \) for short.

SGD update rule is:

\[ w_{t+1} = w_t - \eta_t g_t, \]

where \( g_t = \partial l(w) \).

3. Motivation

Constrained Stochastic Gradient Descent (CSGD)

\[ w^*_{t+1} = \arg \min_{w} \frac{1}{t} \sum_{i=1}^t \frac{1}{2} ||y_i - w^T x_i||^2_2, \]
\[ s.t. \quad w^T x_i = \bar{y}_i. \]

The update rule \( w_{t+1} = P_t(w_t - \eta_t g_t) + r_t \), where \( P_t = I - \frac{x_t x_t^T}{||x_t||^2} \)

\[ r_t = \frac{y_t - \bar{y}_t}{||x_t||^2} x_t. \]

WAIT! Is \( P_t \in \mathbb{R}^{d \times d} \)? YES, but it is rank ONE!

Therefore, we still have a \( O(d) \) time complexity algorithm each iteration:

\[ w_{t+1} = w_t - \eta_t g_t - x_t \left( x_t^T (w_t - \eta_t g_t) \right) / ||x_t||^2 + r_t. \]

4. Algorithm

5. Experiments

Optimization Study

Classification Study

Synthetic dataset

MNIST dataset

---

Remark:

1. performs **consistently better** than SGD in terms of batch optimal.
2. has \( O(\log T) \) regret bound \( R_c(T) = \frac{G^2}{2 \eta} (1 + \log T) \). 3. extensible.