Adaptive Sampling Strategies for Stochastic Optimization

Raghu (Vijaya Raghavendra) Bollapragada[1]
Richard Byrd[2] Jorge Nocedal[1]

[1]Northwestern University
[2]University of Colorado Boulder

January 8, 2018

11^{th} US - Mexico Workshop on Optimization and its Applications
Huatulco, Mexico

Northwestern | McCORMICK SCHOOL OF ENGINEERING

Optimization Problem

$$\min_{x \in \mathbb{R}^d} F(x) = \mathbb{E}_{\zeta}[f(x; \zeta)]$$

Optimization Problem

$$\min_{x \in \mathbb{R}^d} F(x) = \mathbb{E}_\zeta[f(x;\zeta)]$$

Structural Risk Minimization $\quad \min_{x \in \mathbb{R}^d} F(x) = \int f(x;z,y)dP(z,y)$

Optimization Problem

$$\min_{x \in \mathbb{R}^d} F(x) = \mathbb{E}_\zeta[f(x;\zeta)]$$

Structural Risk Minimization $\quad \min_{x \in \mathbb{R}^d} F(x) = \int f(x;z,y)dP(z,y)$

Empirical Risk Minimization $\quad \min_{x \in \mathbb{R}^d} R(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$

Optimization Problem

$$\min_{x \in \mathbb{R}^d} F(x) = \mathbb{E}_\zeta[f(x;\zeta)]$$

Structural Risk Minimization $$\min_{x \in \mathbb{R}^d} F(x) = \int f(x;z,y)dP(z,y)$$

Empirical Risk Minimization $$\min_{x \in \mathbb{R}^d} R(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$$

- Stochastic Gradient is a popular first order method for solving these problems

Optimization Problem

$$\min_{x \in \mathbb{R}^d} F(x) = \mathbb{E}_\zeta[f(x;\zeta)]$$

Structural Risk Minimization $$\min_{x \in \mathbb{R}^d} F(x) = \int f(x;z,y)dP(z,y)$$

Empirical Risk Minimization $$\min_{x \in \mathbb{R}^d} R(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$$

- Stochastic Gradient is a popular first order method for solving these problems
- Many stochastic first order variance reduced methods have been proposed for finite sum problem

  SAG[Schmidt et al. 2016], SAGA[Defazio et al. 2014], SVRG[Johnson and Zhang 2013]

Optimization Problem

$$\min_{x \in \mathbb{R}^d} F(x) = \mathbb{E}_\zeta[f(x;\zeta)]$$

Structural Risk Minimization $$\min_{x \in \mathbb{R}^d} F(x) = \int f(x;z,y)dP(z,y)$$

Empirical Risk Minimization $$\min_{x \in \mathbb{R}^d} R(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$$

- Stochastic Gradient is a popular first order method for solving these problems
- Many stochastic first order variance reduced methods have been proposed for finite sum problem

  SAG[Schmidt et al. 2016], SAGA[Defazio et al. 2014], SVRG[Johnson and Zhang 2013]
- Require either storage or computation of full gradient

Optimization Problem

$$\min_{x \in \mathbb{R}^d} F(x) = \mathbb{E}_\zeta[f(x;\zeta)]$$

Structural Risk Minimization $$\min_{x \in \mathbb{R}^d} F(x) = \int f(x;z,y)dP(z,y)$$

Empirical Risk Minimization $$\min_{x \in \mathbb{R}^d} R(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$$

- Stochastic Gradient is a popular first order method for solving these problems
- Many stochastic first order variance reduced methods have been proposed for finite sum problem

  SAG[Schmidt et al. 2016], SAGA[Defazio et al. 2014], SVRG[Johnson and Zhang 2013]
- Require either storage or computation of full gradient
- Achieve linear convergence for strongly convex functions

Adaptive Sampling Methods

Adaptive Sampling Methods

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k), \qquad \nabla F_{S_k}(x_k) = \frac{1}{|S_k|} \sum_{i \in S_k} \nabla F_i(x_k),$$

where the set $S_k \subset \{1, 2, \ldots\}$ indexes data points (y^i, z^i) drawn at random from the distribution P

Adaptive Sampling Methods

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k), \qquad \nabla F_{S_k}(x_k) = \frac{1}{|S_k|} \sum_{i \in S_k} \nabla F_i(x_k),$$

where the set $S_k \subset \{1, 2, \ldots\}$ indexes data points (y^i, z^i) drawn at random from the distribution P

- Noise in the steps is controlled by sample sizes

Adaptive Sampling Methods

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k), \qquad \nabla F_{S_k}(x_k) = \frac{1}{|S_k|} \sum_{i \in S_k} \nabla F_i(x_k),$$

where the set $S_k \subset \{1, 2, \ldots\}$ indexes data points (y^i, z^i) drawn at random from the distribution P

- Noise in the steps is controlled by sample sizes
- These methods can take advantage of parallel frameworks

Adaptive Sampling Methods

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k), \qquad \nabla F_{S_k}(x_k) = \frac{1}{|S_k|} \sum_{i \in S_k} \nabla F_i(x_k),$$

where the set $S_k \subset \{1, 2, \ldots\}$ indexes data points (y^i, z^i) drawn at random from the distribution P

- Noise in the steps is controlled by sample sizes
- These methods can take advantage of parallel frameworks
- If sample sizes are increased at geometric rate, R-Linear convergence for strongly convex functions

  [Byrd et al. 2012] [Friedlander and Schmidt 2012] [Pasupathy et al. 2015]

Overview

Overview

Norm Test

Norm Test

$$\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\| \leq \theta_n \|\nabla F(x_k)\|, \qquad \text{for some} \quad \theta_n \in [0, 1).$$

Norm Test

$$\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\| \leq \theta_n \|\nabla F(x_k)\|, \qquad \text{for some} \quad \theta_n \in [0, 1).$$

$$\frac{\mathbb{E}[\|\nabla F_i(x_k) - \nabla F(x_k)\|^2]}{|S_k|} \leq \theta_n^2 \|\nabla F(x_k)\|^2$$

Norm Test

$$\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\| \leq \theta_n \|\nabla F(x_k)\|, \qquad \text{for some} \quad \theta_n \in [0, 1).$$

$$\frac{\mathbb{E}[\|\nabla F_i(x_k) - \nabla F(x_k)\|^2]}{|S_k|} \leq \theta_n^2 \|\nabla F(x_k)\|^2$$

- Byrd et al. [2012] proposed this test to control the sample sizes

Norm Test

$$\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\| \leq \theta_n \|\nabla F(x_k)\|, \qquad \text{for some} \quad \theta_n \in [0, 1).$$

$$\frac{\mathbb{E}[\|\nabla F_i(x_k) - \nabla F(x_k)\|^2]}{|S_k|} \leq \theta_n^2 \|\nabla F(x_k)\|^2$$

- Byrd et al. [2012] proposed this test to control the sample sizes
- Cartis and Scheinberg [2016] ensured this condition is satisfied in probability and analyzed global convergence properties

Norm Test

$$\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\| \leq \theta_n \|\nabla F(x_k)\|, \qquad \text{for some} \quad \theta_n \in [0, 1).$$

$$\frac{\mathbb{E}[\|\nabla F_i(x_k) - \nabla F(x_k)\|^2]}{|S_k|} \leq \theta_n^2 \|\nabla F(x_k)\|^2$$

- Byrd et al. [2012] proposed this test to control the sample sizes
- Cartis and Scheinberg [2016] ensured this condition is satisfied in probability and analyzed global convergence properties
- Hashemi et al. [2014] similar test in simulation optimization settings

Norm Test

$$\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\| \leq \theta_n \|\nabla F(x_k)\|, \qquad \text{for some} \quad \theta_n \in [0, 1).$$

$$\frac{\mathbb{E}[\|\nabla F_i(x_k) - \nabla F(x_k)\|^2]}{|S_k|} \leq \theta_n^2 \|\nabla F(x_k)\|^2$$

- Byrd et al. [2012] proposed this test to control the sample sizes
- Cartis and Scheinberg [2016] ensured this condition is satisfied in probability and analyzed global convergence properties
- Hashemi et al. [2014] similar test in simulation optimization settings
- This test is designed to get more than just descent directions

Norm Test

$$\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\| \leq \theta_n \|\nabla F(x_k)\|, \qquad \text{for some} \quad \theta_n \in [0, 1).$$

$$\frac{\mathbb{E}[\|\nabla F_i(x_k) - \nabla F(x_k)\|^2]}{|S_k|} \leq \theta_n^2 \|\nabla F(x_k)\|^2$$

- Byrd et al. [2012] proposed this test to control the sample sizes
- Cartis and Scheinberg [2016] ensured this condition is satisfied in probability and analyzed global convergence properties
- Hashemi et al. [2014] similar test in simulation optimization settings
- This test is designed to get more than just descent directions
- Sample gradients are unnecessarily close to the true gradients

Norm Test

$$\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\| \leq \theta_n \|\nabla F(x_k)\|, \qquad \text{for some} \quad \theta_n \in [0, 1).$$

$$\frac{\mathbb{E}[\|\nabla F_i(x_k) - \nabla F(x_k)\|^2]}{|S_k|} \leq \theta_n^2 \|\nabla F(x_k)\|^2$$

- Byrd et al. [2012] proposed this test to control the sample sizes
- Cartis and Scheinberg [2016] ensured this condition is satisfied in probability and analyzed global convergence properties
- Hashemi et al. [2014] similar test in simulation optimization settings
- This test is designed to get more than just descent directions
- Sample gradients are unnecessarily close to the true gradients
- Sample sizes are increased at much faster rates than the desired rates

Inner Product Test

Inner Product Test

First-Order Descent Condition

$$\nabla F_{S_k}(x_k)^T \nabla F(x_k) > 0$$

Inner Product Test

First-Order Descent Condition

$$\nabla F_{S_k}(x_k)^T \nabla F(x_k) > 0$$

Holds in Expectation

$$\mathbb{E}\left[\nabla F_{S_k}(x_k)^T \nabla F(x_k)\right] = \|\nabla F(x_k)\|^2 > 0$$

Inner Product Test

First-Order Descent Condition

$$\nabla F_{S_k}(x_k)^T \nabla F(x_k) > 0$$

Holds in Expectation

$$\mathbb{E}\left[\nabla F_{S_k}(x_k)^T \nabla F(x_k)\right] = \|\nabla F(x_k)\|^2 > 0$$

For descent condition to hold at most iterations, we impose bounds on the variance

Inner Product Test

First-Order Descent Condition

$$\nabla F_{S_k}(x_k)^T \nabla F(x_k) > 0$$

Holds in Expectation

$$\mathbb{E}\left[\nabla F_{S_k}(x_k)^T \nabla F(x_k)\right] = \|\nabla F(x_k)\|^2 > 0$$

For descent condition to hold at most iterations, we impose bounds on the variance

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4, \; \theta_{ip} \in [0, 1)$$

Inner Product Test

First-Order Descent Condition

$$\nabla F_{S_k}(x_k)^T \nabla F(x_k) > 0$$

Holds in Expectation

$$\mathbb{E}\left[\nabla F_{S_k}(x_k)^T \nabla F(x_k)\right] = \|\nabla F(x_k)\|^2 > 0$$

For descent condition to hold at most iterations, we impose bounds on the variance

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4, \; \theta_{ip} \in [0, 1)$$

- Test is designed to achieve descent directions sufficiently often

Inner Product Test

First-Order Descent Condition

$$\nabla F_{S_k}(x_k)^T \nabla F(x_k) > 0$$

Holds in Expectation

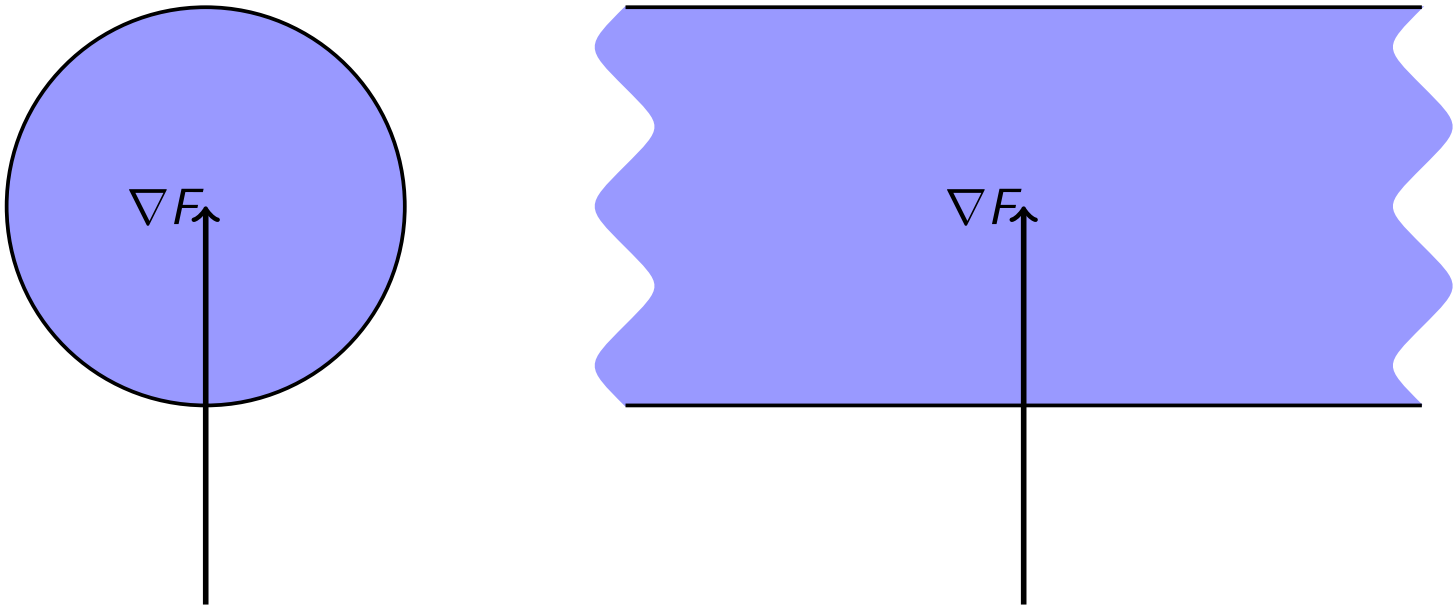
$$\mathbb{E}\left[\nabla F_{S_k}(x_k)^T \nabla F(x_k)\right] = \|\nabla F(x_k)\|^2 > 0$$

For descent condition to hold at most iterations, we impose bounds on the variance

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4, \; \theta_{ip} \in [0, 1)$$

- Test is designed to achieve descent directions sufficiently often
- Samples sizes required to satisfy this condition are smaller than those required for norm condition

Comparison



(a)
Norm Test
$\|g(x) - \nabla F(x)\| \leq \theta_n \|\nabla F(x)\|$

(b)
Inner Product Test
$\left|g(x)^T \nabla F(x) - \|\nabla F(x)\|^2\right| \leq \theta_{ip} \|\nabla F(x)\|^2$

Figure: Given a gradient ∇F the shaded areas denote the set of vectors satisfying the deterministic (a) Norm test (b) Inner Product test

Comparison

Lemma

Let $|S_{ip}|$, $|S_n|$ represent the minimum number of samples required to satisfy the inner product test and norm test at any given iterate x and any given $\theta_{ip} = \theta_n < 1$. Then we have

$$\frac{|S_{ip}|}{|S_n|} = \beta(x) \leq 1,$$

where

$$\beta(x) = \frac{\mathbb{E}[\|\nabla F_i(x)\|^2 \cos^2(\gamma_i)] - \|\nabla F(x)\|^2}{\mathbb{E}[\|\nabla F_i(x)\|^2] - \|\nabla F(x)\|^2},$$

and γ_i is the angle made by $\nabla F_i(x)$ with $\nabla F(x)$.

Test Approximation

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4, \; \theta_{ip} \in [0, 1)$$

Test Approximation

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4, \; \theta_{ip} \in [0, 1)$$

- Computing true gradient is expensive

Test Approximation

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4, \; \theta_{ip} \in [0, 1)$$

- Computing true gradient is expensive
- Approximate population variance with sample variance and true gradient with sampled gradient

Test Approximation

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4,\ \theta_{ip} \in [0,1)$$

- Computing true gradient is expensive
- Approximate population variance with sample variance and true gradient with sampled gradient

$$\frac{\mathrm{Var}_{i \in S_k}(\nabla F_i(x_k)^T \nabla F_{S_k}(x_k))}{|S_k|} \leq \theta_{ip}^2 \|\nabla F_{S_k}(x_k)\|^4,$$

Test Approximation

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4,\ \theta_{ip} \in [0,1)$$

- Computing true gradient is expensive
- Approximate population variance with sample variance and true gradient with sampled gradient

$$\frac{\mathrm{Var}_{i \in S_k}(\nabla F_i(x_k)^T \nabla F_{S_k}(x_k))}{|S_k|} \leq \theta_{ip}^2 \|\nabla F_{S_k}(x_k)\|^4,$$

$$\mathrm{Var}_{i \in S_k}\left(\nabla F_i(x_k)^T \nabla F_{S_k}(x_k)\right) = \frac{1}{|S_k| - 1} \sum_{i \in S_k} \left(\nabla F_i(x_k)^T \nabla F_{S_k}(x_k) - \|\nabla F_{S_k}(x_k)\|^2\right)^2$$

Test Approximation

$$\frac{\mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]}{|S_k|} \leq \theta_{ip}^2 \|\nabla F(x_k)\|^4, \; \theta_{ip} \in [0, 1)$$

- Computing true gradient is expensive
- Approximate population variance with sample variance and true gradient with sampled gradient

$$\frac{\mathrm{Var}_{i \in S_k}(\nabla F_i(x_k)^T \nabla F_{S_k}(x_k))}{|S_k|} \leq \theta_{ip}^2 \|\nabla F_{S_k}(x_k)\|^4,$$

$$\mathrm{Var}_{i \in S_k}\left(\nabla F_i(x_k)^T \nabla F_{S_k}(x_k)\right) = \frac{1}{|S_k| - 1} \sum_{i \in S_k} \left(\nabla F_i(x_k)^T \nabla F_{S_k}(x_k) - \|\nabla F_{S_k}(x_k)\|^2\right)^2$$

- Whenever condition is not satisfied, increase sample size to satisfy the condition

Backup

Overview

Orthogonal Test

Orthogonal Test

- Although the Inner Product test is practical, convergence cannot be established because it allows sample gradients that are arbitrarily long relative to $\|\nabla F(x_k)\|$

Orthogonal Test

- Although the Inner Product test is practical, convergence cannot be established because it allows sample gradients that are arbitrarily long relative to $\|\nabla F(x_k)\|$
- No restriction on near orthogonality of sample gradient and gradient

Orthogonal Test

- Although the Inner Product test is practical, convergence cannot be established because it allows sample gradients that are arbitrarily long relative to $\|\nabla F(x_k)\|$
- No restriction on near orthogonality of sample gradient and gradient
- Component of sample gradient orthogonal to gradient is 0 in expectation

Orthogonal Test

- Although the Inner Product test is practical, convergence cannot be established because it allows sample gradients that are arbitrarily long relative to $\|\nabla F(x_k)\|$
- No restriction on near orthogonality of sample gradient and gradient
- Component of sample gradient orthogonal to gradient is 0 in expectation
- We control the variance in the orthogonal components of sampled gradients

$$\frac{\mathbb{E}\left[\left\|\nabla F_i(x_k) - \frac{\nabla F_i(x_k)^T \nabla F(x_k)}{\|\nabla F(x_k)\|^2}\nabla F(x_k)\right\|^2\right]}{|S_k|} \leq \nu^2 \|\nabla F(x_k)\|^2,\ \nu > 0$$

Linear Convergence

Linear Convergence

Theorem

Suppose that F is twice continuously differentiable and that there exist constants $0 < \mu \leq L$ such that

$$\mu I \preceq \nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Linear Convergence

Theorem

Suppose that F is twice continuously differentiable and that there exist constants $0 < \mu \leq L$ such that

$$\mu I \preceq \nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$.

Linear Convergence

Theorem

Suppose that F is twice continuously differentiable and that there exist constants $0 < \mu \leq L$ such that

$$\mu I \preceq \nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$. Then, if the steplength satisfies

$$\alpha_k = \alpha = \frac{1}{(1 + \theta_{ip}^2 + \nu^2)L},$$

Linear Convergence

Theorem

Suppose that F is twice continuously differentiable and that there exist constants $0 < \mu \leq L$ such that

$$\mu I \preceq \nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$. Then, if the steplength satisfies

$$\alpha_k = \alpha = \frac{1}{(1 + \theta_{ip}^2 + \nu^2)L},$$

we have that

$$\mathbb{E}[F(w_k) - F(w^*)] \leq \rho^k (F(x_0) - F(x^*)),$$

where

$$\rho = 1 - \frac{\mu}{L} \frac{1}{(1 + \theta_{ip}^2 + \nu^2)}$$

Convex Non-Convex

Overview

Step-Length Selection

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k)$$

Step-Length Selection

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k)$$

- Stochastic gradient is employed with diminishing stepsizes

Step-Length Selection

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k)$$

- Stochastic gradient is employed with diminishing stepsizes
- Exact line search can be performed to determine the stepsize but it is too expensive

Step-Length Selection

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k)$$

- Stochastic gradient is employed with diminishing stepsizes
- Exact line search can be performed to determine the stepsize but it is too expensive
- Constant stepsize can be employed but one needs to know the Lipschitz constant of the problem

$$\alpha_k = 1/L$$

Step-Length Selection

$$x_{k+1} = x_k - \alpha_k \nabla F_{S_k}(x_k)$$

- Stochastic gradient is employed with diminishing stepsizes
- Exact line search can be performed to determine the stepsize but it is too expensive
- Constant stepsize can be employed but one needs to know the Lipschitz constant of the problem

$$\alpha_k = 1/L$$

- We propose to estimate the Lipschitz constant as we proceed, resulting in adaptive stepsizes

Step-Length Selection

Algorithm 1 Estimating Lipschitz constant

Input: $L_{k-1} > 0$, some $\eta > 1$

1: Compute parameter $\zeta_k > 1$
2: Set $L_k = L_{k-1}/\zeta_k$; ▷ Decrease the Lipschitz constant
3: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
4: **while** $F_{new} > F_{s_k}(x_k) - \frac{1}{2L_k}\|\nabla F_{s_k}(x_k)\|^2$ **do** ▷ sufficient decrease
5: Set $L_k = \eta L_{k-1}$ ▷ Increase the Lipschitz constant
6: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
7: **end while**

Step-Length Selection

Algorithm 1 Estimating Lipschitz constant

Input: $L_{k-1} > 0$, some $\eta > 1$

1: Compute parameter $\zeta_k > 1$
2: Set $L_k = L_{k-1}/\zeta_k$; ▷ Decrease the Lipschitz constant
3: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
4: **while** $F_{new} > F_{s_k}(x_k) - \frac{1}{2L_k}\|\nabla F_{s_k}(x_k)\|^2$ **do** ▷ sufficient decrease
5: Set $L_k = \eta L_{k-1}$ ▷ Increase the Lipschitz constant
6: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
7: **end while**

- Beck and Teboulle [2009] proposed this algorithm to estimate Lipschitz constant in deterministic settings

Step-Length Selection

Algorithm 1 Estimating Lipschitz constant

Input: $L_{k-1} > 0$, some $\eta > 1$

1: Compute parameter $\zeta_k > 1$
2: Set $L_k = L_{k-1}/\zeta_k$; ▷ Decrease the Lipschitz constant
3: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
4: **while** $F_{new} > F_{s_k}(x_k) - \frac{1}{2L_k}\|\nabla F_{s_k}(x_k)\|^2$ **do** ▷ sufficient decrease
5: Set $L_k = \eta L_{k-1}$ ▷ Increase the Lipschitz constant
6: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
7: **end while**

- Beck and Teboulle [2009] proposed this algorithm to estimate Lipschitz constant in deterministic settings
- Scmidt et al. [2016] adapted this algorithm to stochastic algorithms such as SAG

Step-Length Selection

Algorithm 1 Estimating Lipschitz constant

Input: $L_{k-1} > 0$, some $\eta > 1$

1: Compute parameter $\zeta_k > 1$
2: Set $L_k = L_{k-1}/\zeta_k$; ▷ Decrease the Lipschitz constant
3: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
4: **while** $F_{new} > F_{s_k}(x_k) - \frac{1}{2L_k}\|\nabla F_{s_k}(x_k)\|^2$ **do** ▷ sufficient decrease
5: Set $L_k = \eta L_{k-1}$ ▷ Increase the Lipschitz constant
6: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
7: **end while**

- Similar to Line-Search on sampled functions with memory

Step-Length Selection

Algorithm 1 Estimating Lipschitz constant

Input: $L_{k-1} > 0$, some $\eta > 1$

1: Compute parameter $\zeta_k > 1$
2: Set $L_k = L_{k-1}/\zeta_k$; ▷ Decrease the Lipschitz constant
3: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
4: **while** $F_{new} > F_{s_k}(x_k) - \frac{1}{2L_k}\|\nabla F_{s_k}(x_k)\|^2$ **do** ▷ sufficient decrease
5: Set $L_k = \eta L_{k-1}$ ▷ Increase the Lipschitz constant
6: Compute $F_{new} = F_{s_k}\left(x_k - \frac{1}{L_k}\nabla F_{s_k}(x_k)\right)$
7: **end while**

- Similar to Line-Search on sampled functions with memory
- Only needs access to sampled function values

Aggressive Steps Heuristic

Aggressive Steps Heuristic

It is well known and easy to show that

$$\mathbb{E}[F(x_{k+1})] - F(x_k) \leq -\alpha_k \|F(x_k)\|^2 + \frac{\alpha_k^2 L}{2} \mathbb{E}[\|\nabla F_{S_k}(x_k)\|^2]$$

Aggressive Steps Heuristic

It is well known and easy to show that

$$\mathbb{E}[F(x_{k+1})] - F(x_k) \leq -\alpha_k \|F(x_k)\|^2 + \frac{\alpha_k^2 L}{2}\mathbb{E}[\|\nabla F_{S_k}(x_k)\|^2]$$

Thus, we can obtain a decrease in the true objective, in expectation, if

$$\frac{L\alpha_k^2}{2}\left(\mathbb{E}[\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\|^2] + \|\nabla F(x_k)\|^2\right) \leq \alpha_k \|\nabla F(x_k)\|^2$$

Aggressive Steps Heuristic

$$\frac{L\alpha_k^2}{2}\left(\mathbb{E}[\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\|^2] + \|\nabla F(x_k)\|^2\right) \leq \alpha_k \|\nabla F(x_k)\|^2$$

Aggressive Steps Heuristic

$$\frac{L\alpha_k^2}{2}\left(\mathbb{E}[\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\|^2] + \|\nabla F(x_k)\|^2\right) \leq \alpha_k \|\nabla F(x_k)\|^2$$

Using, $\alpha_k = 1/L_k$, assuming $L_{k-1} \geq L$, and sample approximations

$$L_k \geq \frac{L_{k-1}}{2}\left(\frac{\text{Var}\left(\nabla F_i(x_k)\right)}{|S_k|\|\nabla F_{S_k}(x_k)\|^2} + 1\right),$$

where $\text{Var}\left(\nabla F_i(x_k)\right) = \frac{1}{|S_k|-1}\sum_{i\in S_k}\|\nabla F_i(x_k) - \nabla F_{S_k}(x_k)\|^2$.

Aggressive Steps Heuristic

$$\frac{L\alpha_k^2}{2}\left(\mathbb{E}[\|\nabla F_{S_k}(x_k) - \nabla F(x_k)\|^2] + \|\nabla F(x_k)\|^2\right) \leq \alpha_k \|\nabla F(x_k)\|^2$$

Using, $\alpha_k = 1/L_k$, assuming $L_{k-1} \geq L$, and sample approximations

$$L_k \geq \frac{L_{k-1}}{2}\left(\frac{\text{Var}\left(\nabla F_i(x_k)\right)}{|S_k|\|\nabla F_{S_k}(x_k)\|^2} + 1\right),$$

where $\text{Var}\left(\nabla F_i(x_k)\right) = \frac{1}{|S_k|-1}\sum_{i\in S_k}\|\nabla F_i(x_k) - \nabla F_{S_k}(x_k)\|^2$.
Therefore,

$$\zeta_k = \max\left(1, \frac{2}{\frac{\text{Var}(\nabla F_i(x_k))}{|S_k|\|\nabla F_{S_k}(x_k)\|^2} + 1}\right)$$

Parameter Selection

Parameter Selection

$$\frac{\nabla F_{S_k}(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2}{\left(\frac{\sigma}{\sqrt{|S_k|}}\right)} \sim \mathcal{N}(0,1),$$

where $\sigma^2 = \mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]$ is the true variance.

Parameter Selection

$$\frac{\nabla F_{S_k}(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2}{\left(\frac{\sigma}{\sqrt{|S_k|}}\right)} \sim \mathcal{N}(0,1),$$

where $\sigma^2 = \mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]$ is the true variance.

- Parameter θ_{ip} is directly proportional to probability of getting a descent direction

Parameter Selection

$$\frac{\nabla F_{S_k}(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2}{\left(\frac{\sigma}{\sqrt{|S_k|}}\right)} \sim \mathcal{N}(0,1),$$

where $\sigma^2 = \mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]$ is the true variance.

- Parameter θ_{ip} is directly proportional to probability of getting a descent direction
- $\theta_{ip} = 0.7$ corresponds to 0.9 probability

Parameter Selection

$$\frac{\nabla F_{S_k}(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2}{\left(\frac{\sigma}{\sqrt{|S_k|}}\right)} \sim \mathcal{N}(0,1),$$

where $\sigma^2 = \mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]$ is the true variance.

- Parameter θ_{ip} is directly proportional to probability of getting a descent direction
- $\theta_{ip} = 0.7$ corresponds to 0.9 probability
- $\theta_{ip} = 0.9$ works well in practice

Parameter Selection

$$\frac{\nabla F_{S_k}(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2}{\left(\frac{\sigma}{\sqrt{|S_k|}}\right)} \sim \mathcal{N}(0,1),$$

where $\sigma^2 = \mathbb{E}\left[\left(\nabla F_i(x_k)^T \nabla F(x_k) - \|\nabla F(x_k)\|^2\right)^2\right]$ is the true variance.

- Parameter θ_{ip} is directly proportional to probability of getting a descent direction
- $\theta_{ip} = 0.7$ corresponds to 0.9 probability
- $\theta_{ip} = 0.9$ works well in practice
- Orthogonal test is seldom active in practice and we choose $\nu = \tan(80^o) = 5.84$ for all problems

Overview

Results: Constant Step-Length Strategy

$$R(x) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-z^i x^T y^i)) + \frac{\lambda}{2}\|x\|^2$$

Results: Constant Step-Length Strategy

$$R(x) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-z^i x^T y^i)) + \frac{\lambda}{2}\|x\|^2$$

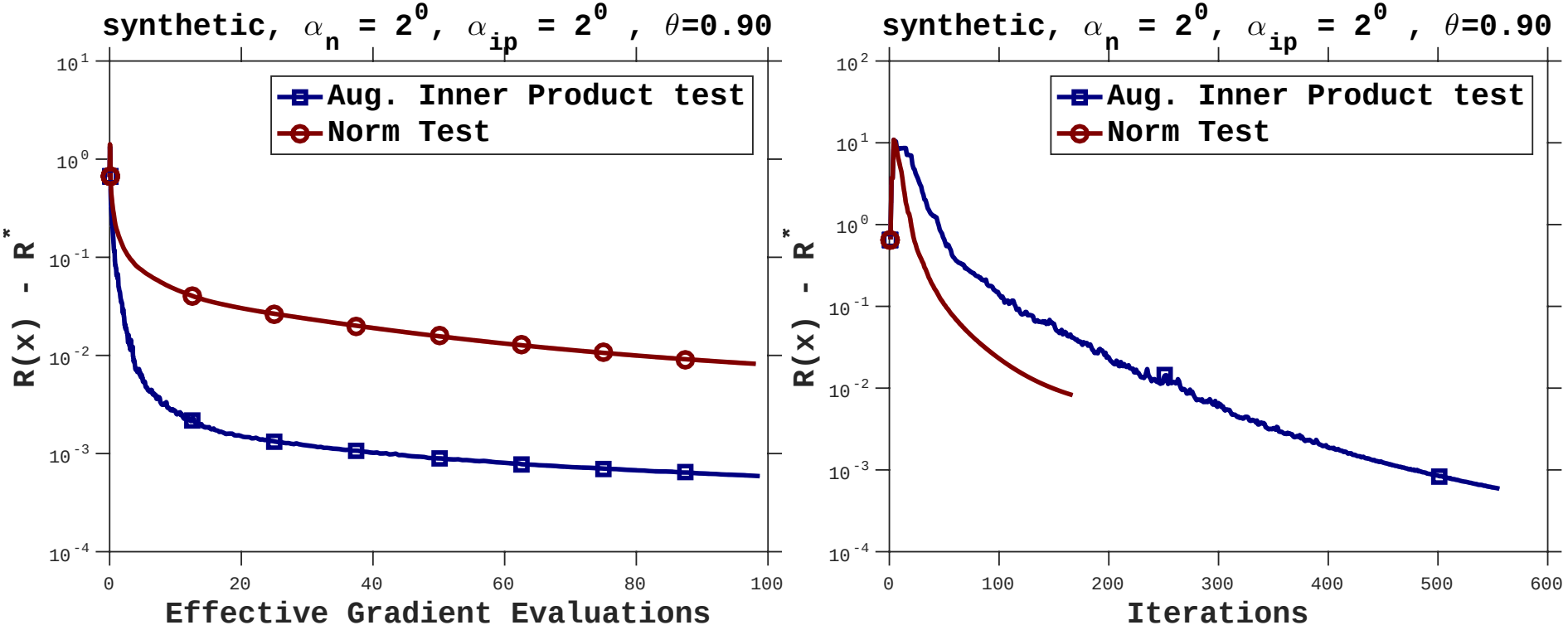


Figure: Norm Test vs. Inner Product Test. Synthetic Dataset ($n = 7000$)

Results: Constant Step-Length Strategy

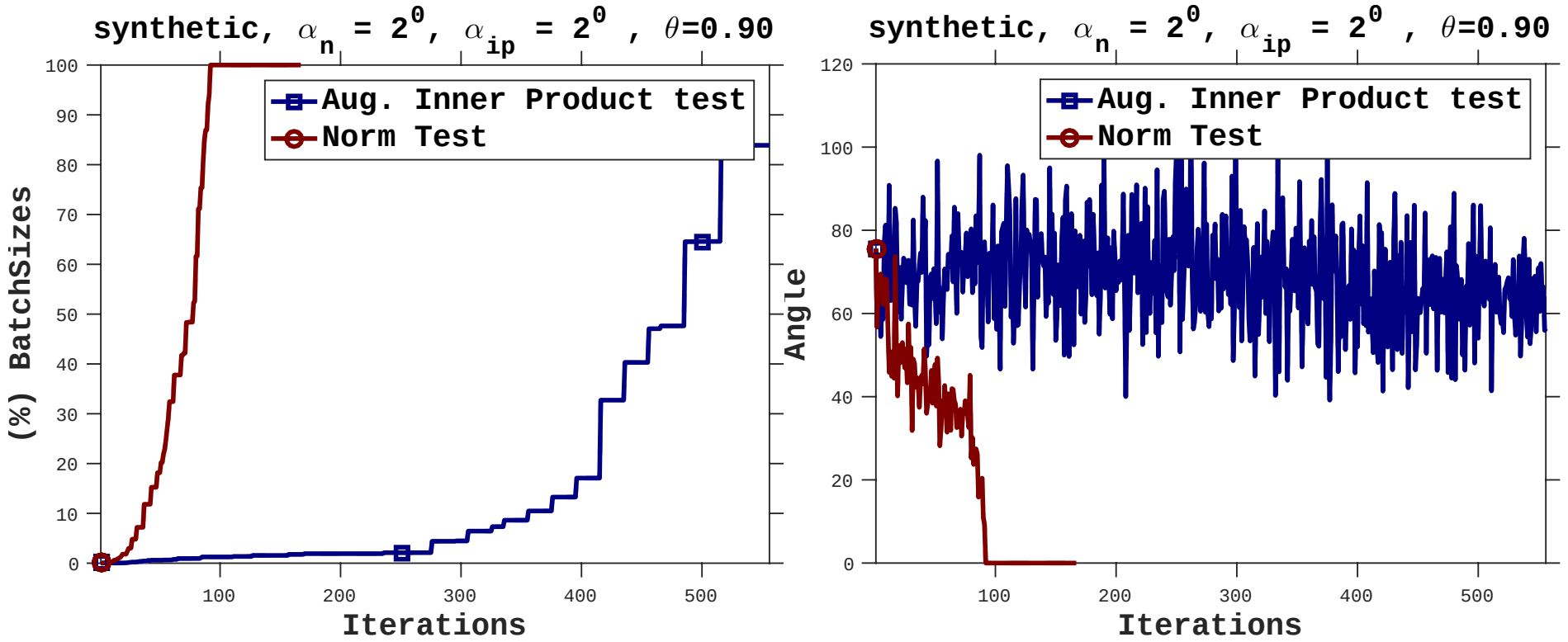
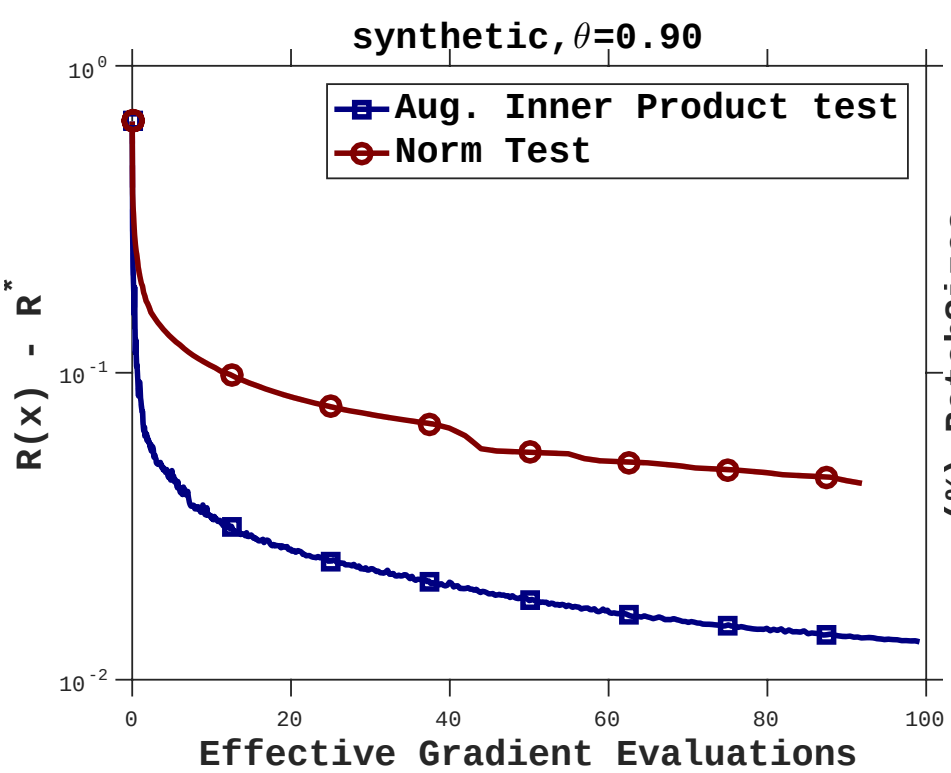


Figure: Synthetic Dataset ($n = 7000$)

Other Datasets

Results: Adaptive Step-Length Strategy



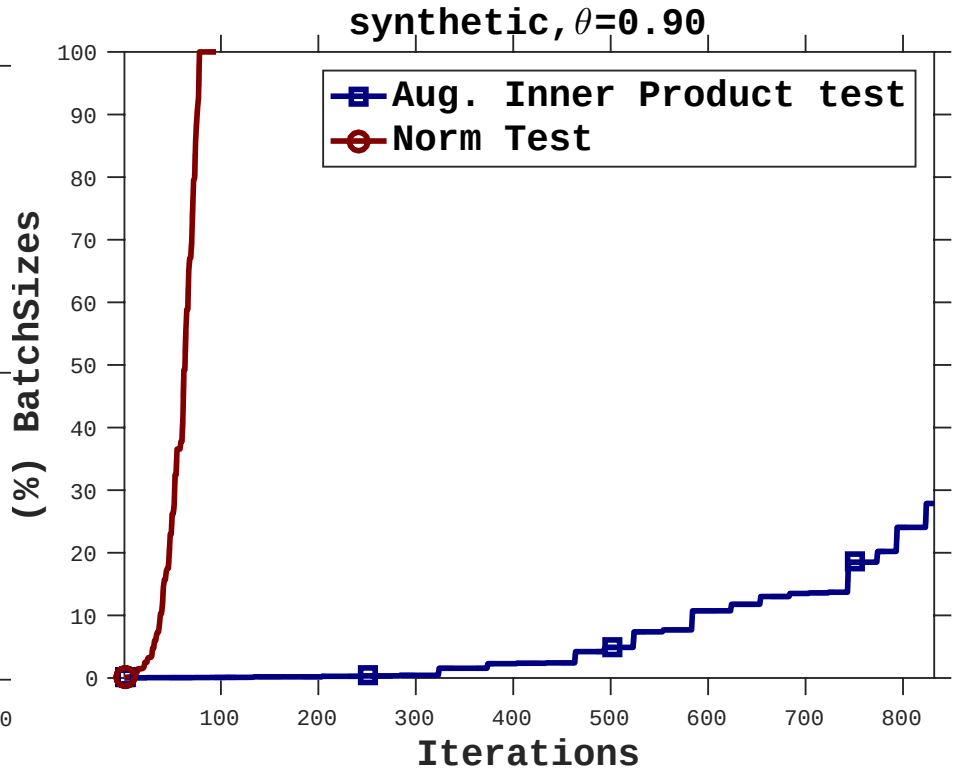


Figure: Synthetic Dataset ($n = 7000$)

Results: Adaptive Step-Length Strategy

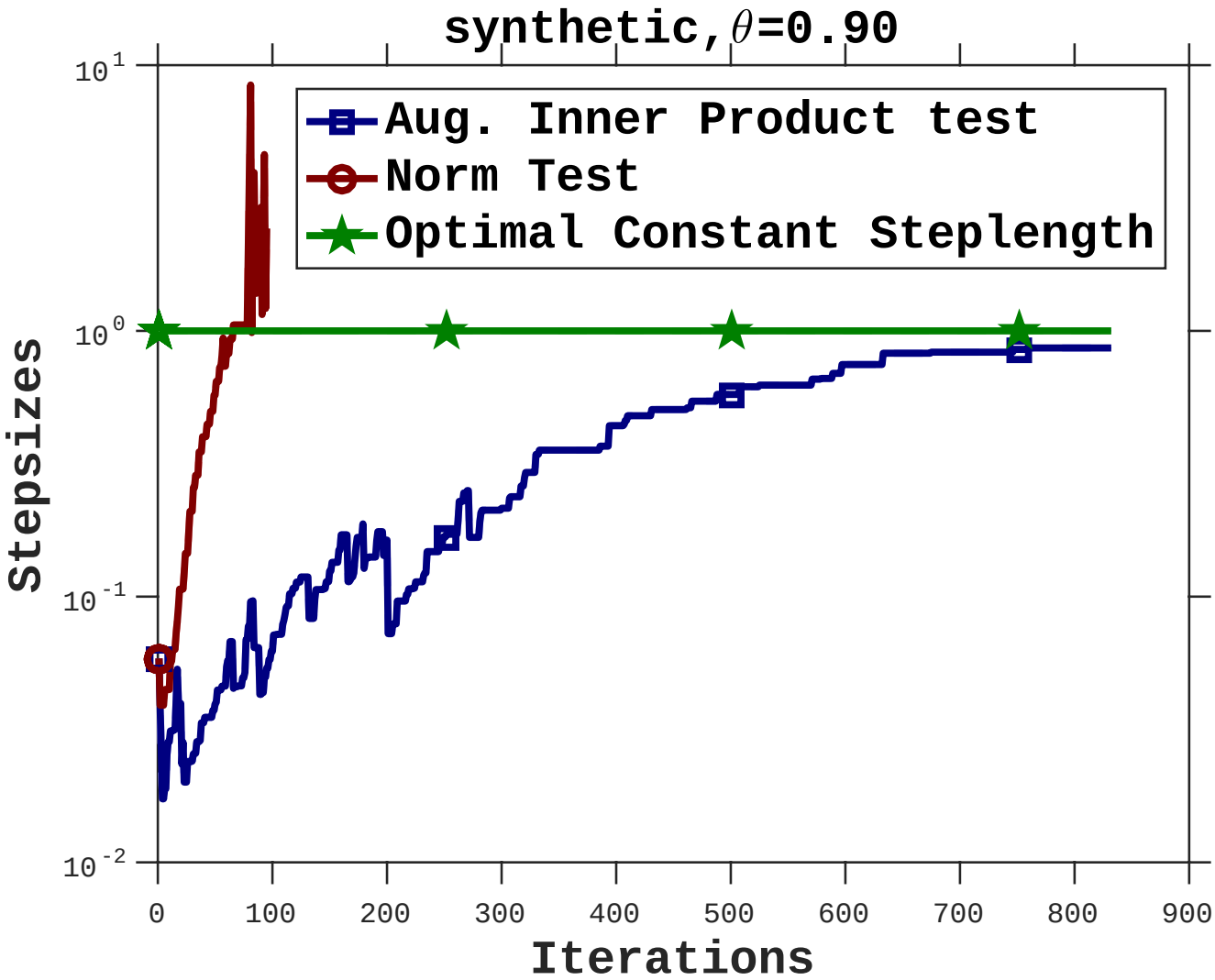


Figure: Synthetic Dataset ($n = 7000$)

Other Datasets

Overview

Summary

Summary

- Adaptive sampling methods are alternate methods for noise reduction

Summary

- Adaptive sampling methods are alternate methods for noise reduction
- These methods can lead to speed ups when implemented in parallel environments

Summary

- Adaptive sampling methods are alternate methods for noise reduction
- These methods can lead to speed ups when implemented in parallel environments
- We propose a practical inner product test which is better at controlling the sample sizes than the existing norm test

Summary

- Adaptive sampling methods are alternate methods for noise reduction
- These methods can lead to speed ups when implemented in parallel environments
- We propose a practical inner product test which is better at controlling the sample sizes than the existing norm test
- These methods can use adaptive stepsizes and second-order information can be incorporated

Summary

- Adaptive sampling methods are alternate methods for noise reduction
- These methods can lead to speed ups when implemented in parallel environments
- We propose a practical inner product test which is better at controlling the sample sizes than the existing norm test
- These methods can use adaptive stepsizes and second-order information can be incorporated
- Currently working on practical sampling tests to control sample sizes in stochastic quasi-Newton methods

Questions???

Thank You

Appendix

Backup Mechanism

Backup Mechanism

- Sample approximations are not accurate when samples are very small (say 1, 5 , 10)

Backup Mechanism

- Sample approximations are not accurate when samples are very small (say 1, 5 , 10)
- Our tests may not be accurate in controlling the sample sizes in such situations

Backup Mechanism

- Sample approximations are not accurate when samples are very small (say 1, 5 , 10)
- Our tests may not be accurate in controlling the sample sizes in such situations
- Need more accurate approximations in such scenarios

Backup Mechanism

- Sample approximations are not accurate when samples are very small (say 1, 5 , 10)
- Our tests may not be accurate in controlling the sample sizes in such situations
- Need more accurate approximations in such scenarios

$$g_{avg} \stackrel{\text{def}}{=} \frac{1}{r} \sum_{j=k-r+1}^{k} \nabla F_{S_j}(x_j)$$

Backup Mechanism

- Sample approximations are not accurate when samples are very small (say 1, 5 , 10)
- Our tests may not be accurate in controlling the sample sizes in such situations
- Need more accurate approximations in such scenarios

$$g_{avg} \stackrel{\text{def}}{=} \frac{1}{r} \sum_{j=k-r+1}^{k} \nabla F_{S_j}(x_j)$$

- r should be chosen such that the iterates in the summation are close enough and there are enough samples for g_{avg} to be accurate ($r = 10$).
- If $\|g_{avg}\| < \gamma \|\nabla F_{S_k}(x_k)\|$, for some $\gamma \in (0, 1)$ then we use g_{avg} instead of $\nabla F_{S_k}(x_k)$ in the tests.

Main Presentation

Convex Functions

Convex Functions

Theorem

(General Convex Objective.) Suppose that F is twice continuously differentiable and convex, and that there exists a constant $L > 0$ such that

$$\nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Convex Functions

Theorem

(General Convex Objective.) Suppose that F is twice continuously differentiable and convex, and that there exists a constant $L > 0$ such that

$$\nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$.

Convex Functions

Theorem

(General Convex Objective.) Suppose that F is twice continuously differentiable and convex, and that there exists a constant $L > 0$ such that

$$\nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$. Then, if the steplength satisfies

$$\alpha_k = \alpha < \frac{1}{(1 + \theta_{ip}^2 + \nu^2)L},$$

Convex Functions

Theorem

(General Convex Objective.) Suppose that F is twice continuously differentiable and convex, and that there exists a constant $L > 0$ such that

$$\nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$. Then, if the steplength satisfies

$$\alpha_k = \alpha < \frac{1}{(1 + \theta_{ip}^2 + \nu^2)L},$$

we have for any positive integer T,

$$\min_{0 \le k \le T-1} \mathbb{E}\left[F\left(x_k\right)\right] - F^* \le \frac{1}{2\alpha c T} \|x_0 - x^*\|^2,$$

where the constant $c > 0$ is given by $c = 1 - L\alpha(1 + \theta_{ip}^2 + \nu^2)$.

Main-Presentation

Non-Convex Functions

Non-Convex Functions

Theorem

(Nonconvex Objective.) Suppose that F is twice continuously differentiable and bounded below, and that there exist a constant $L > 0$ such that

$$\nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Non-Convex Functions

Theorem

(Nonconvex Objective.) Suppose that F is twice continuously differentiable and bounded below, and that there exist a constant $L > 0$ such that

$$\nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$.

Non-Convex Functions

Theorem

(Nonconvex Objective.) Suppose that F is twice continuously differentiable and bounded below, and that there exist a constant $L > 0$ such that

$$\nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$. Then, if the steplength satisfies

$$\alpha_k = \alpha \leq \frac{1}{(1 + \theta_{ip}^2 + \nu^2)L},$$

Non-Convex Functions

Theorem

(Nonconvex Objective.) Suppose that F is twice continuously differentiable and bounded below, and that there exist a constant $L > 0$ such that

$$\nabla^2 F(x) \preceq LI, \quad \forall x \in \mathbb{R}^d.$$

Let $\{x_k\}$ be the iterates generated by subsampled gradient method with any x_0, where $|S_k|$ is chosen such that inner product test and orthogonal test are satisfied at each iteration for any given $\theta_{ip} > 0$ and $\nu > 0$. Then, if the steplength satisfies

$$\alpha_k = \alpha \leq \frac{1}{(1 + \theta_{ip}^2 + \nu^2)L},$$

we have that

$$\lim_{k \to \infty} \mathbb{E}[\|\nabla F(x_k)\|^2] \to 0.$$

Moreover, for any positive integer T we have that

$$\min_{0 \leq k \leq T-1} \mathbb{E}[\|\nabla F(x_k)\|^2] \leq \frac{2}{\alpha T}(F(x_0) - F_{min}).$$

Main-Presentation

Results: Constant Step-Length Strategy

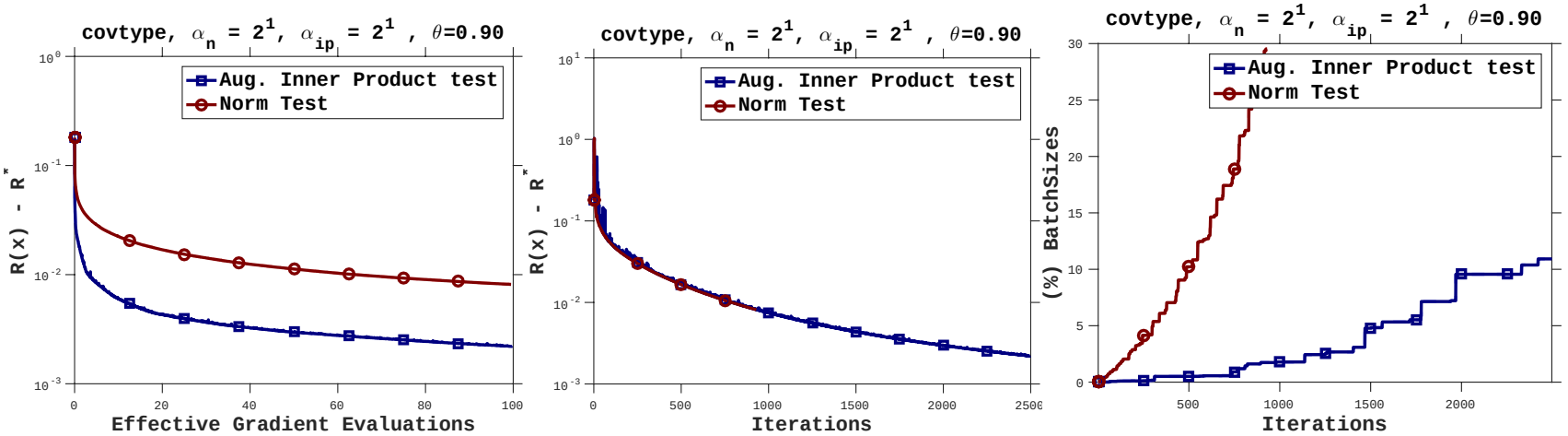


Figure: Covertype Dataset (n = 581012)

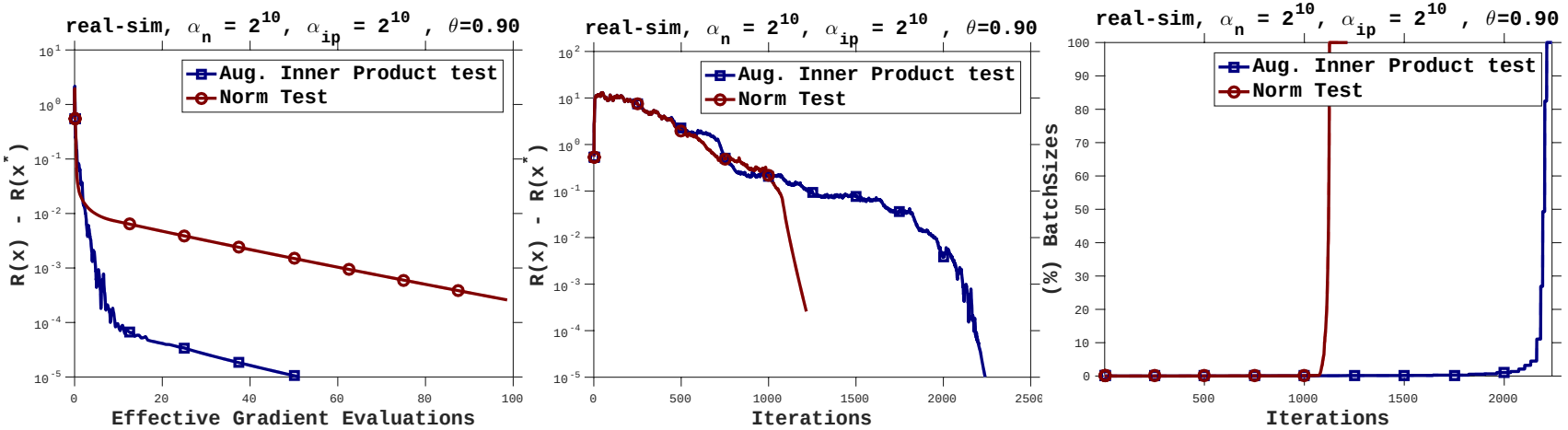


Figure: Real-Sim Dataset (n = 65078)

Results: Constant Step-Length Strategy

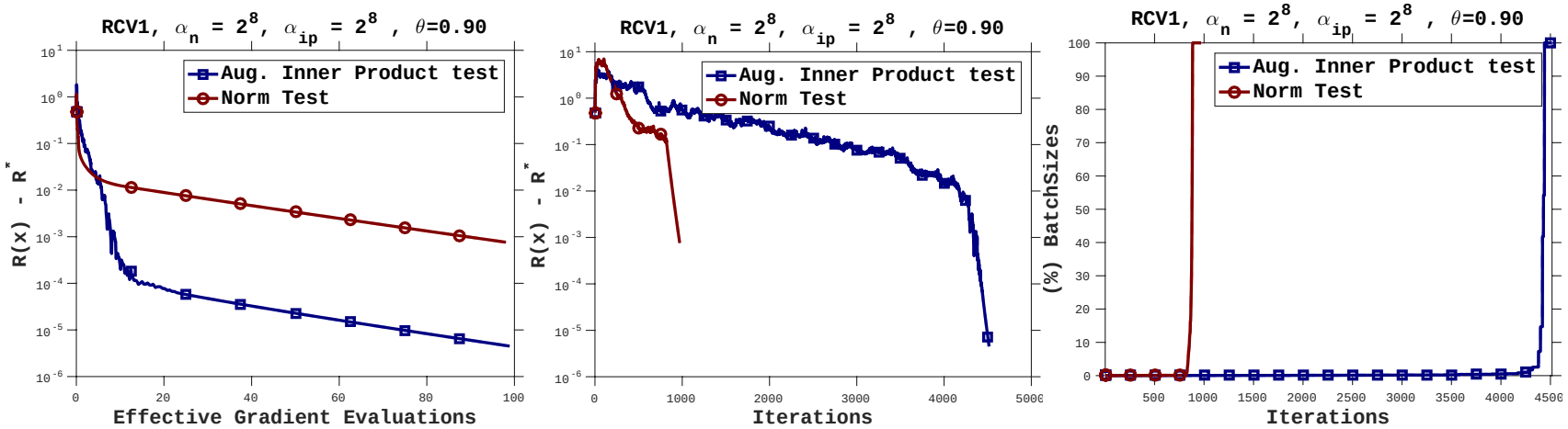


Figure: RCV1 Dataset (n = 20242)

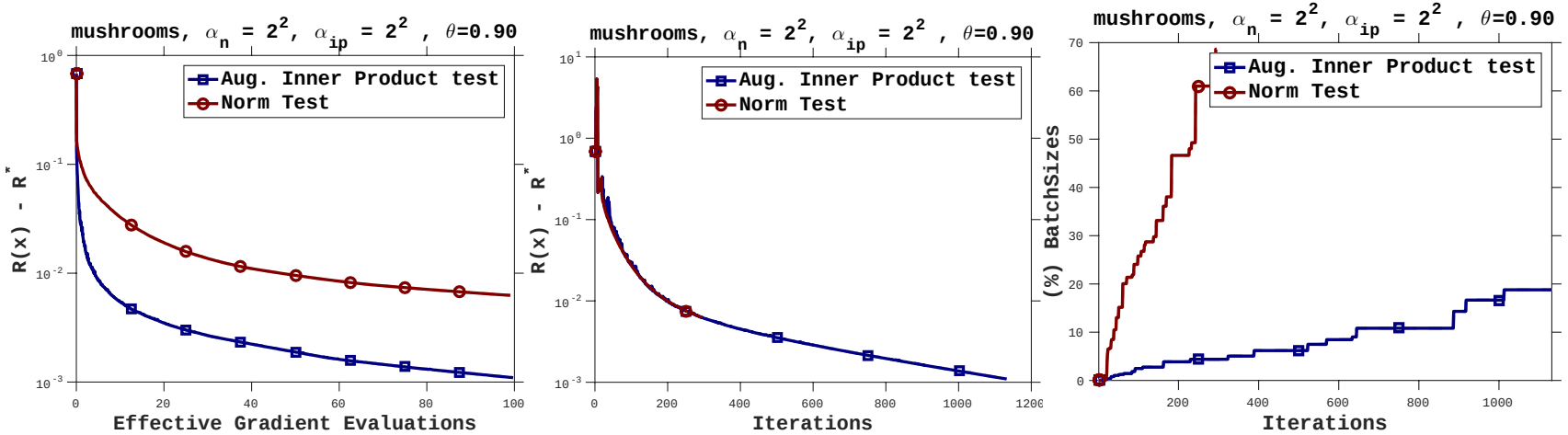


Figure: Mushrooms Dataset (n = 8124)

Results: Constant Step-Length Strategy

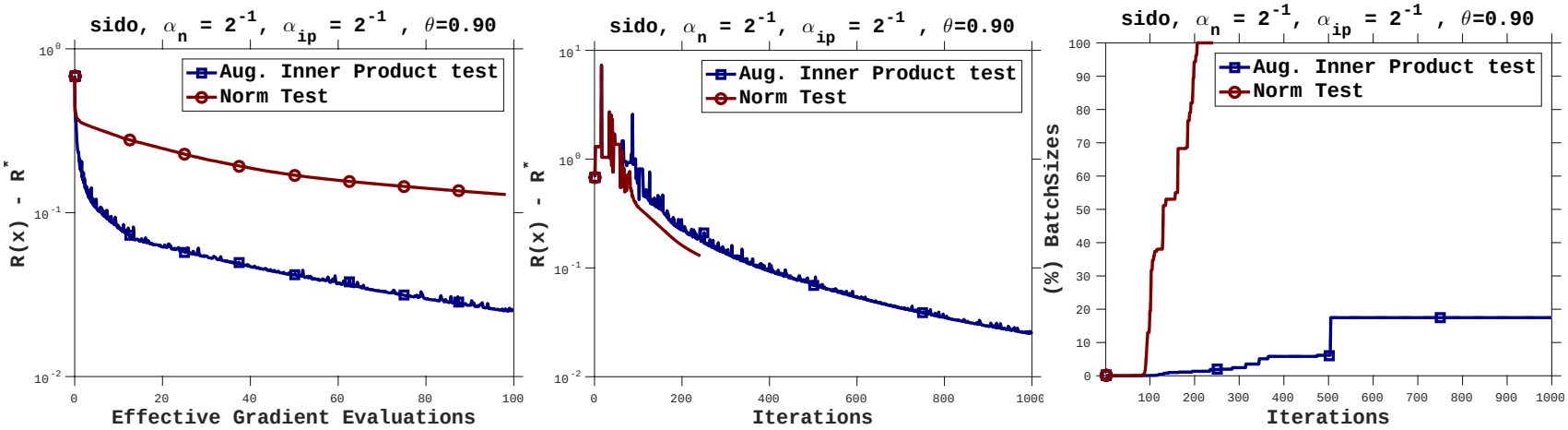


Figure: Sido Dataset (n = 12678)

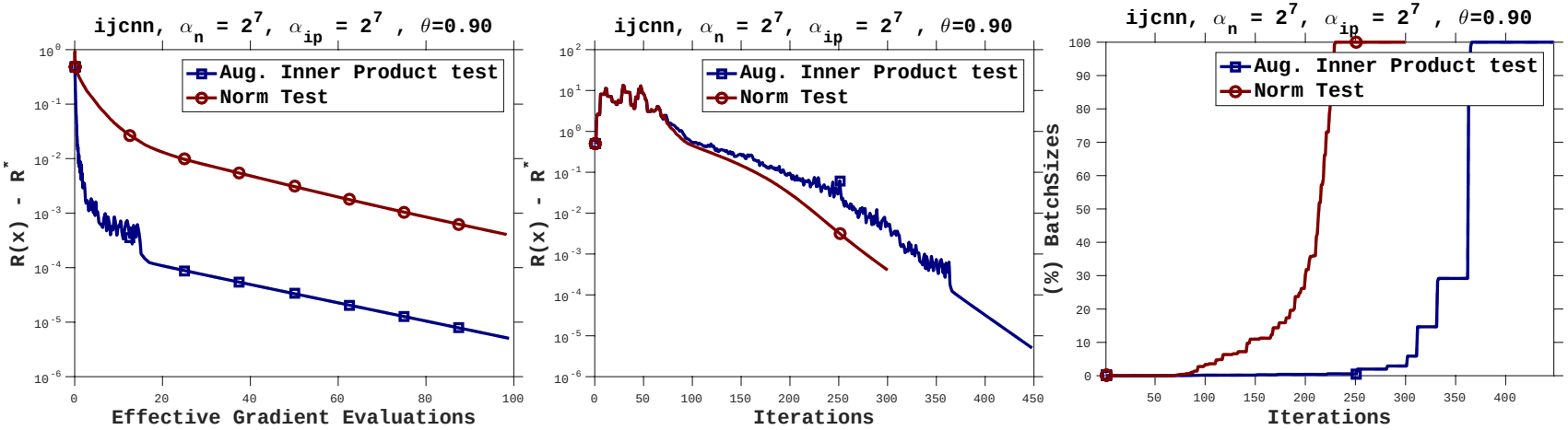


Figure: Ijcnn Dataset (n = 35000)

Results: Constant Step-Length Strategy

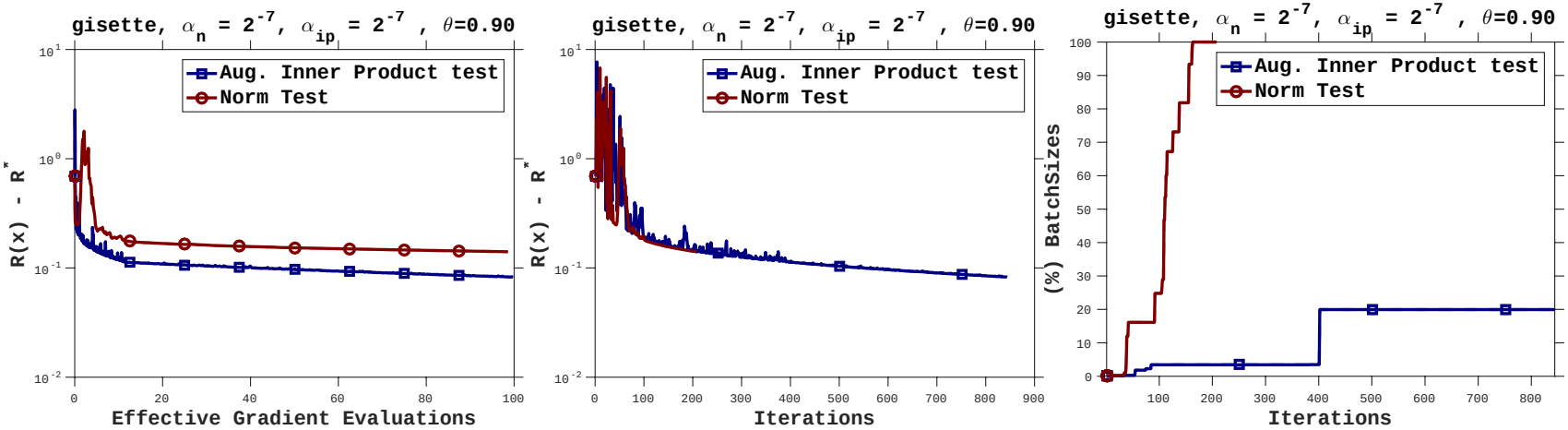


Figure: Gisette Dataset (n = 6000)

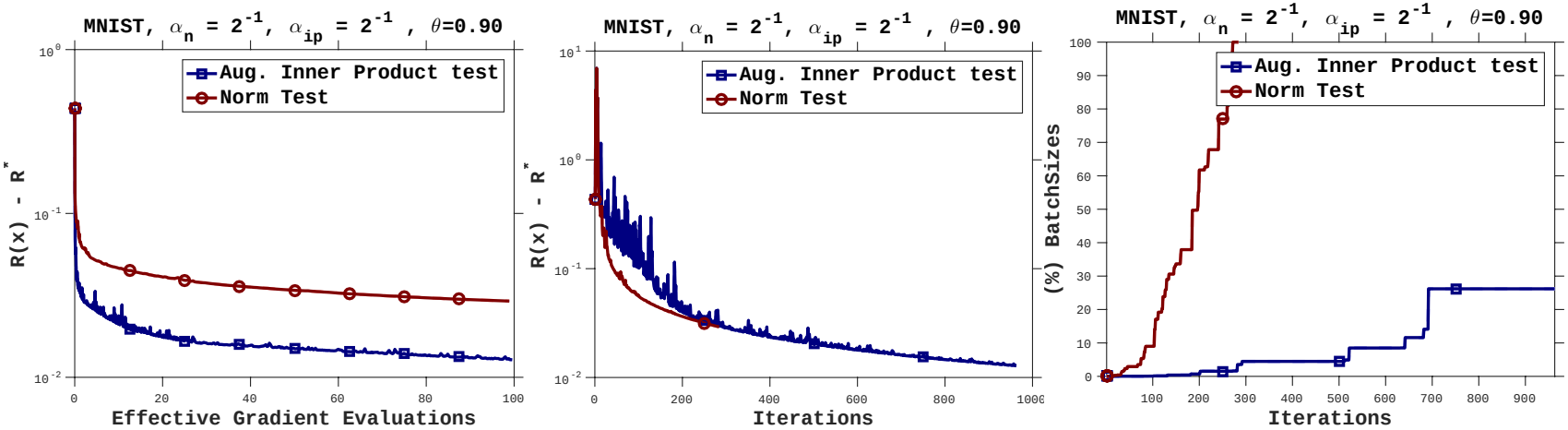


Figure: MNIST Dataset (n = 60000)

Results: Adaptive Step-Length Strategy

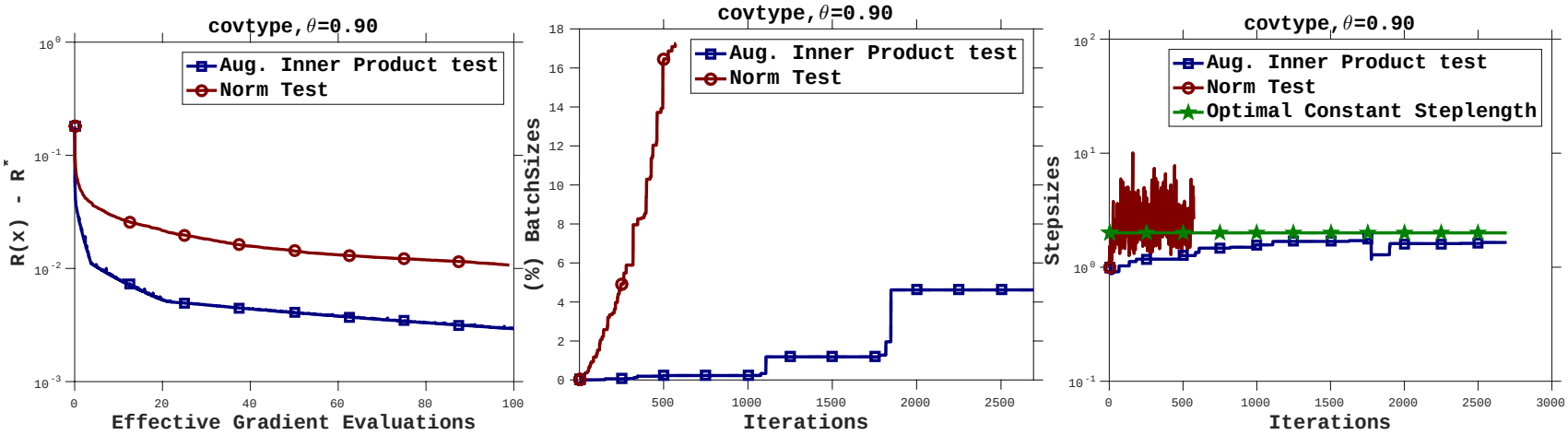


Figure: Covertype Dataset (n = 581012)

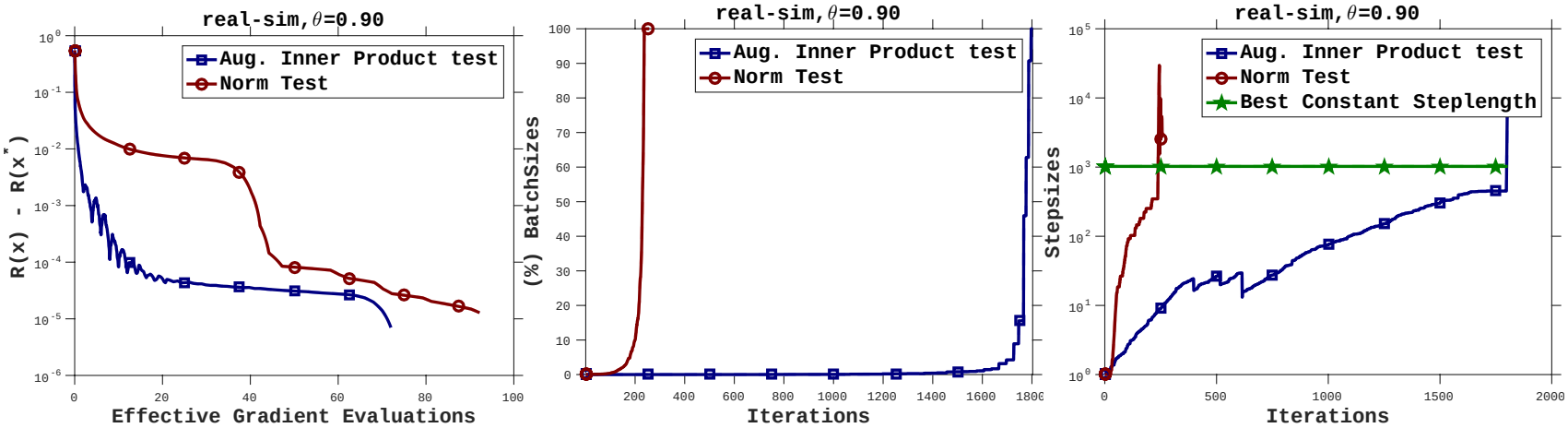


Figure: Real-Sim Dataset (n = 65078)

Results: Adaptive Step-Length Strategy

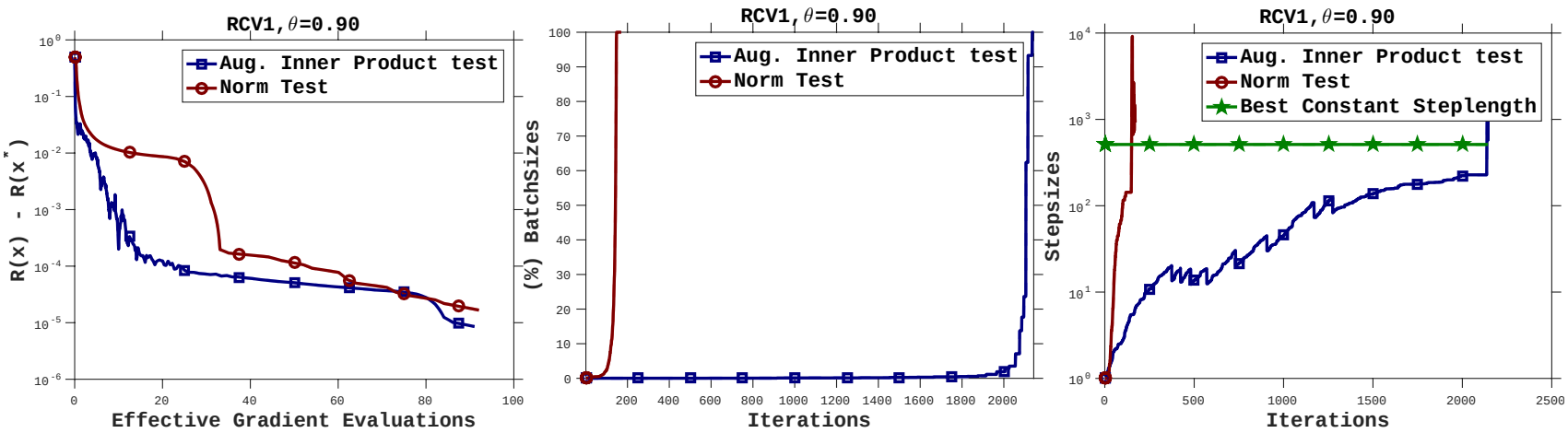


Figure: RCV1 Dataset (n = 20242)

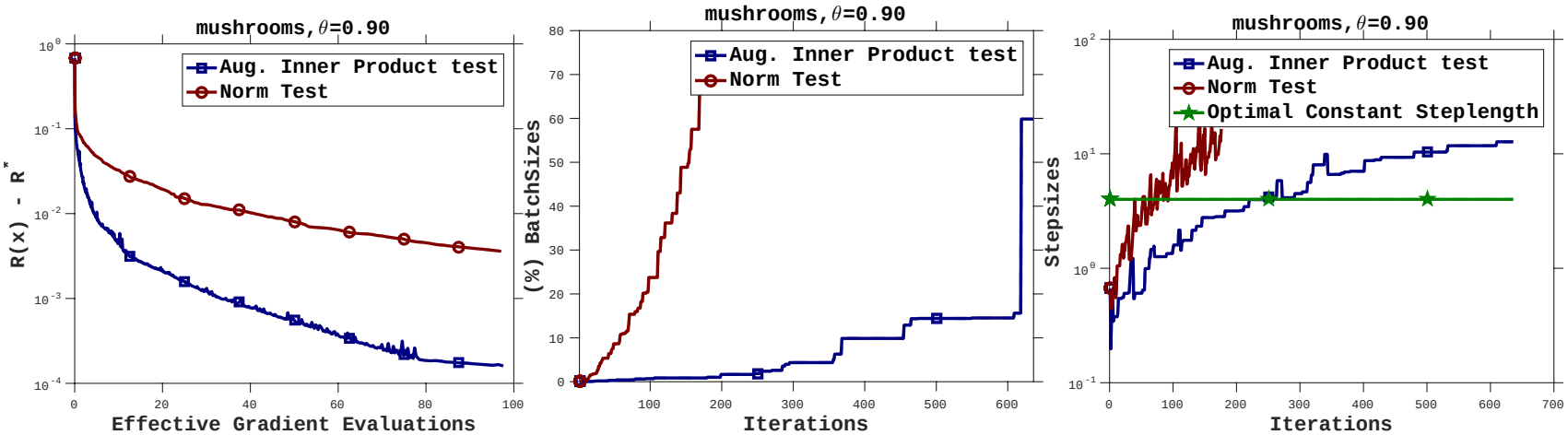


Figure: Mushrooms Dataset (n = 8124)

Results: Adaptive Step-Length Strategy

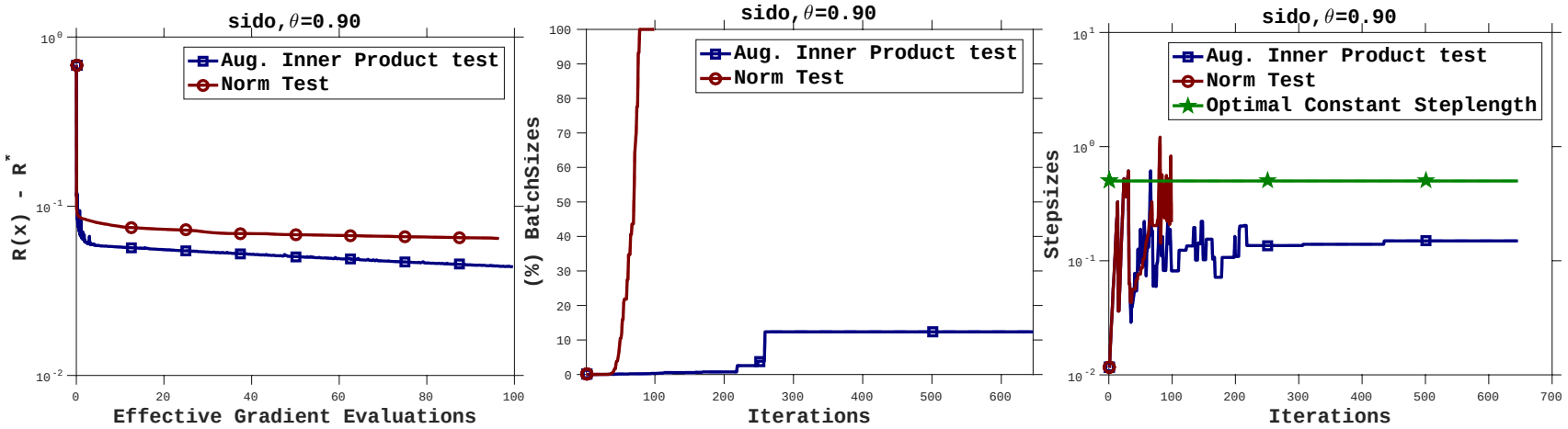


Figure: Sido Dataset (n = 12678)

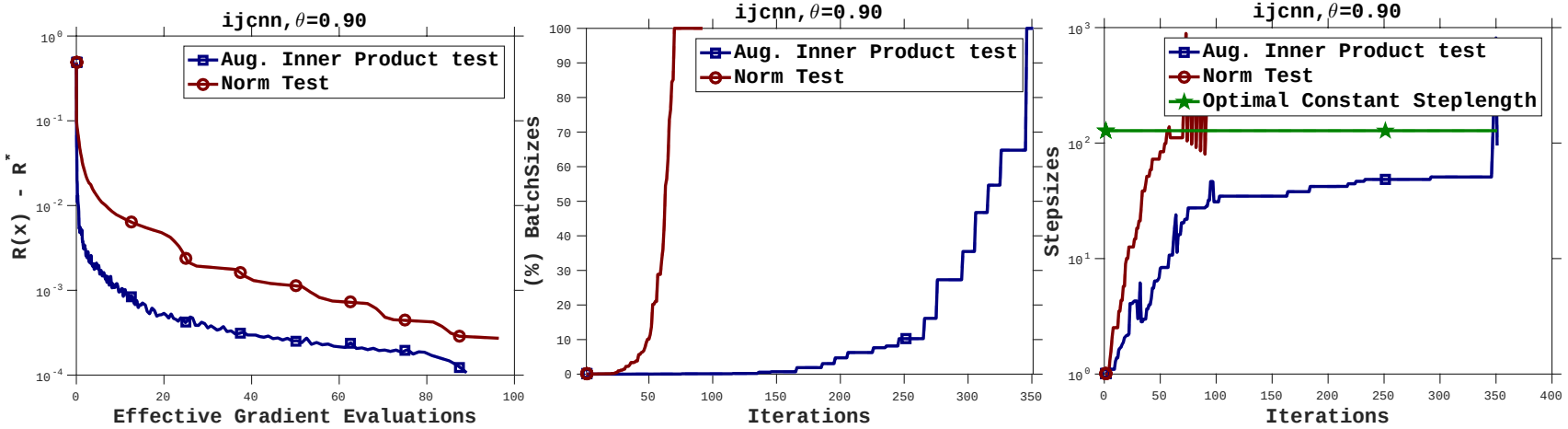


Figure: Ijcnn Dataset (n = 35000)

Results: Adaptive Step-Length Strategy

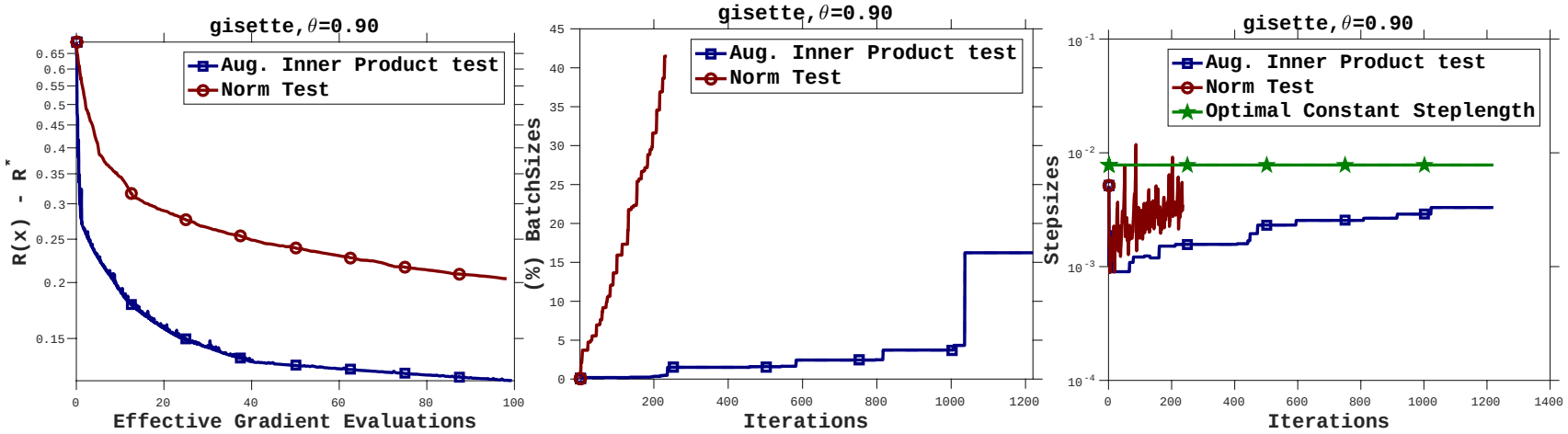


Figure: Gisette Dataset (n = 6000)

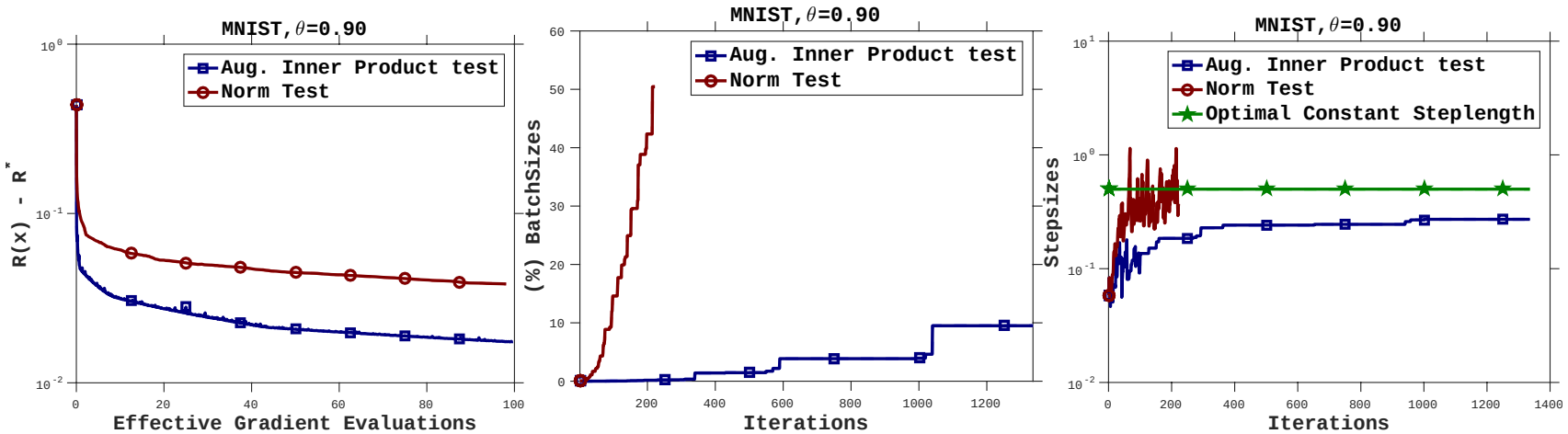


Figure: MNIST Dataset (n = 60000)