

# S01 supplemental methods

ORCID: 0000-0000-0000-1234 Email: author@example.com DOI: 10.5281/zenodo.12345678

November 13, 2025

## Contents

1	Supplemental Methods	1
1.1	S1.1 Extended Algorithm Variants	1
1.1.1	S1.1.1 Stochastic Variant	1
1.1.2	S1.1.2 Mini-Batch Variant	2
1.2	S1.2 Detailed Convergence Analysis	2
1.2.1	S1.2.1 Strong Convexity Assumptions	2
1.2.2	S1.2.2 Lipschitz Continuity	2
1.3	S1.3 Additional Theoretical Results	2
1.3.1	S1.3.1 Worst-Case Complexity Bounds	2
1.3.2	S1.3.2 Expected Convergence for Stochastic Variants	3
1.4	S1.4 Implementation Considerations	3
1.4.1	S1.4.1 Numerical Stability	3
1.4.2	S1.4.2 Initialization Strategies	3
1.5	S1.5 Extended Mathematical Framework	3
1.5.1	S1.5.1 Generalized Objective Function	3
1.5.2	S1.5.2 Adaptive Weight Selection	3
1.6	S1.6 Convergence Diagnostics	4
1.6.1	S1.6.1 Diagnostic Criteria	4
1.6.2	S1.6.2 Failure Detection	4
1.7	S1.7 Parameter Sensitivity	4

## 1 Supplemental Methods

This section provides detailed methodological information that supplements Section ??.

### 1.1 S1.1 Extended Algorithm Variants

#### 1.1.1 S1.1.1 Stochastic Variant

For large-scale problems, we developed a stochastic variant of our algorithm:

$$x_{k+1} = x_k - f_{i_k}(x_k) + \eta_k(x_k - x_{k-1}) \quad (1.1)$$

where  $i_k$  is a randomly sampled index from  $\{1, \dots, n\}$  at iteration  $k$ .

**Convergence Analysis:** Under appropriate sampling strategies, this variant achieves  $O(1/k)$  convergence rate for non-strongly convex problems, following the analysis in [? ?].

### 1.1.2 S1.1.2 Mini-Batch Variant

To balance between computational efficiency and convergence speed:

$$x_{k+1} = x_k - \frac{1}{|B_k|} \sum_{i \in B_k} f_i(x_k) + \gamma_k(x_k - x_{k-1}) \quad (1.2)$$

where  $B_k \subseteq \{1, \dots, n\}$  is a mini-batch of size  $|B_k| = b$ .

## 1.2 S1.2 Detailed Convergence Analysis

### 1.2.1 S1.2.1 Strong Convexity Assumptions

We assume the objective function  $f$  satisfies:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\gamma}{2} \|y - x\|^2, \quad x, y \in X \quad (1.3)$$

where  $\gamma > 0$  is the strong convexity parameter.

### 1.2.2 S1.2.2 Lipschitz Continuity

The gradient is Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad x, y \in X \quad (1.4)$$

The condition number  $\kappa = L/\gamma$  determines the convergence rate:  $\kappa = 1/(1/\gamma)$ , as established in [? ?].

## 1.3 S1.3 Additional Theoretical Results

### 1.3.1 S1.3.1 Worst-Case Complexity Bounds

**Theorem S1:** Under the assumptions of Lipschitz continuity and strong convexity, the algorithm requires at most  $O((1/\gamma))$  iterations to achieve  $\epsilon$ -accuracy.

**Proof:** From the convergence rate (??), we have:

$$\|x_k - x^*\| \leq C^k \sqrt{k} \frac{(C/\gamma)}{(1/\gamma)} = O((1/\gamma)) \quad (1.5)$$

since  $(1/\gamma) \leq 1/\epsilon$  for small  $1/\gamma$ .  $\square$

### 1.3.2 S1.3.2 Expected Convergence for Stochastic Variants

For the stochastic variant (1.1):

$$E[x_k | x^2] \leq \frac{C}{k} + \sigma^2 \quad (1.6)$$

where  $\sigma^2$  is the variance of the stochastic gradient estimates.

## 1.4 S1.4 Implementation Considerations

### 1.4.1 S1.4.1 Numerical Stability

To ensure numerical stability, we implement the following safeguards:

1. Gradient clipping:  $f(x_k) \in (1, /f(x_k))f(x_k)$
2. Step size bounds:  $\alpha_k$
3. Momentum bounds:  $0 \leq \alpha_k \leq 1$

### 1.4.2 S1.4.2 Initialization Strategies

We tested three initialization strategies:

1. Random:  $x_0 \sim N(0, I)$
2. Warm start:  $x_0 = \text{solution from simpler problem}$
3. Problem-specific:  $x_0 = \text{domain knowledge-based initialization}$

Results show that warm start initialization reduces iterations by approximately 30% for related problem instances.

## 1.5 S1.5 Extended Mathematical Framework

### 1.5.1 S1.5.1 Generalized Objective Function

The framework extends to more general objectives:

$$f(x) = \sum_{i=1}^n w_i(x) + \sum_{j=1}^m R_j(x) + \sum_{k=1}^p C_k(x) \quad (1.7)$$

where:  
-  $w_i(x)$ : Data fitting terms  
-  $R_j(x)$ : Regularization terms (e.g.,  $l_1$ ,  $l_2$ , elastic net)  
-  $C_k(x)$ : Constraint terms (penalty or barrier functions)

### 1.5.2 S1.5.2 Adaptive Weight Selection

Weights  $w_i$  can be adapted during optimization:

$$w_i^{(k+1)} = w_i^{(k)} \left( \frac{|i(x_k)|}{|(x_k)|} \right) \quad (1.8)$$

This reweighting scheme gives more emphasis to terms that are harder to optimize.

## 1.6 S1.6 Convergence Diagnostics

### 1.6.1 S1.6.1 Diagnostic Criteria

We monitor the following quantities for convergence:

1. Gradient norm:  $f(x_k) < g$
2. Step size:  $x_{k+1} - x_k < x$
3. Function improvement:  $|f(x_{k+1}) - f(x_k)| < f$
4. Relative improvement:  $|f(x_{k+1}) - f(x_k)| / |f(x_k)| < r$

All four criteria must be satisfied for declared convergence.

### 1.6.2 S1.6.2 Failure Detection

Algorithm failure is detected if:

1. Maximum iterations exceeded
2. Step size becomes too small ( $\kappa < \epsilon$ )
3. NaN or Inf values encountered
4. Objective function increases for consecutive iterations

## 1.7 S1.7 Parameter Sensitivity

Detailed sensitivity analysis for each parameter:

Parameter	Nominal	Range	Impact on Performance
$\alpha$	0.01	[0.001, 0.1]	High (±30%)
	0.9	[0.5, 0.99]	Medium (±15%)
	0.001	[0, 0.01]	Low (±5%)

Table 1. Parameter sensitivity analysis results

The learning rate  $\alpha$  has the strongest impact on convergence speed, while regularization primarily affects the final solution quality rather than convergence dynamics.