Supplemental Methods

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

S1.1 Extended Algorithm Variants

S1.1.1 Stochastic Variant

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

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k) + \beta_k(x_k - x_{k-1}) \tag{1}$$

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

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

S1.1.2 Mini-Batch Variant

To balance between computational efficiency and convergence speed:

$$x_{k+1} = x_k - \alpha_k \frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k) + \beta_k(x_k - x_{k-1}) \tag{2}$$

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

S1.2 Detailed Convergence Analysis

S1.2.1 Strong Convexity Assumptions

We assume the objective function f satisfies:

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{X} \tag{3}$$

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

S1.2.2 Lipschitz Continuity

The gradient is Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{X} \tag{4}$$

The condition number $\kappa = L/\mu$ determines the convergence rate: $\rho = \sqrt{1 - 1/\kappa}$, as established in [**?**, **?**].

S1.3 Additional Theoretical Results

S1.3.1 Worst-Case Complexity Bounds

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

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

$$\|x_k - x^*\| \leq C\rho^k \leq \epsilon \Rightarrow k \geq \frac{\log(C/\epsilon)}{\log(1/\rho)} = O(\kappa \log(1/\epsilon)) \tag{5}$$

since $\log(1/\rho) \approx 1/\kappa$ for small $1/\kappa$. $\square$

S1.3.2 Expected Convergence for Stochastic Variants

For the stochastic variant (1):

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{C}{k} + \sigma^2 \tag{6}$$

where σ^2 is the variance of the stochastic gradient estimates.

S1.4 Implementation Considerations

S1.4.1 Numerical Stability

To ensure numerical stability, we implement the following safeguards:

1. **Gradient clipping**: $\nabla f(x_k) \leftarrow \min(1, \theta/\|\nabla f(x_k)\|)\nabla f(x_k)$
2. **Step size bounds**: $\alpha_{\min} \leq \alpha_k \leq \alpha_{\max}$
3. **Momentum bounds**: $0 \leq \beta_k \leq \beta_{\max} < 1$

S1.4.2 Initialization Strategies

We tested three initialization strategies:

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

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

S1.5 Extended Mathematical Framework

S1.5.1 Generalized Objective Function

The framework extends to more general objectives:

$$f(x) = \sum_{i=1}^{n} w_i \phi_i(x) + \sum_{j=1}^{m} \lambda_j R_j(x) + \sum_{k=1}^{p} \gamma_k C_k(x) \tag{7}$$

where: - $\phi_i(x)$: 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)

S1.5.2 Adaptive Weight Selection

Weights w_i can be adapted during optimization:

$$w_i^{(k+1)} = w_i^{(k)} \cdot \exp\left(-\gamma \frac{|\phi_i(x_k)|}{|\phi(x_k)|}\right) \tag{8}$$

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

S1.6 Convergence Diagnostics

S1.6.1 Diagnostic Criteria

We monitor the following quantities for convergence:

1. **Gradient norm**: $\|\nabla f(x_k)\| < \epsilon_g$
2. **Step size**: $\|x_{k+1} - x_k\| < \epsilon_x$
3. **Function improvement**: $|f(x_{k+1}) - f(x_k)| < \epsilon_f$
4. **Relative improvement**: $|f(x_{k+1}) - f(x_k)|/|f(x_k)| < \epsilon_r$

All four criteria must be satisfied for declared convergence.

S1.6.2 Failure Detection

Algorithm failure is detected if:

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

S1.7 Parameter Sensitivity

Detailed sensitivity analysis for each parameter:

The learning rate α_0 has the strongest impact on convergence speed, while regularization λ primarily affects the final solution quality rather than convergence dynamics.

Parameter	Nominal	Range	Impact on Performance
α_0	0.01	[0.001, 0.1]	High (±30%)
β	0.9	[0.5, 0.99]	Medium (±15%)
λ	0.001	[0, 0.01]	Low (±5%)

Table 1: Parameter sensitivity analysis results