Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization

Tyler B. Johnson and Carlos Guestrin, University of Washington

Motivation

Approaches to scaling optimization:

- Stochastic methods
- Parallelism
- Active sets — Approach of this work

Our focus is constrained convex optimization:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & h_j(\mathbf{x}) \le 0 \quad j = 1, \ldots, m\,. \end{aligned} \tag{P1}$$

Given solution $\mathbf{x}^\star$, define $\mathcal{C}^\star = \{h_j : h_j(\mathbf{x}^\star) = 0\}$. Then $\mathbf{x}^\star$ also solves

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & h_j(\mathbf{x}) \le 0 \quad h_j \in \mathcal{C}^\star\,. \end{aligned} \tag{P2}$$

Main idea:

To solve (P1) efficiently, solve problems similar to (P2)!

Active Set Algorithms

Since $\mathcal{C}^\star$ is unknown, we use active sets:

Until converged **do**
- $\mathcal{C} \leftarrow$ Prioritized set of constraints
- $\mathbf{x} \leftarrow$ Minimizer of f subject to constraints in $\mathcal{C}$

(Common active sets set-up)

Traditionally very fast in practice but difficult to analyze.

Blitz Algorithm

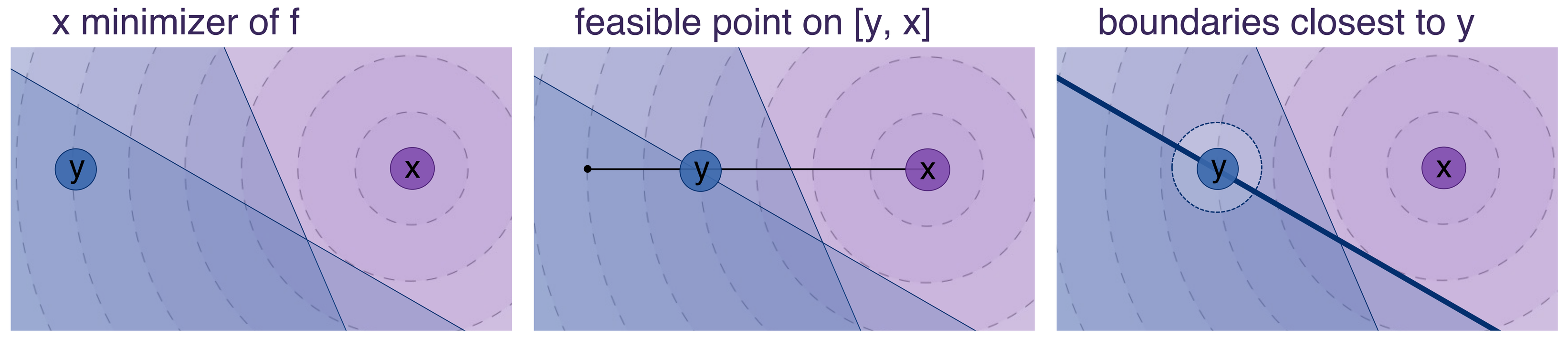
Blitz is a fast, novel active set algorithm with useful guarantees.

Algorithm 1 Blitz

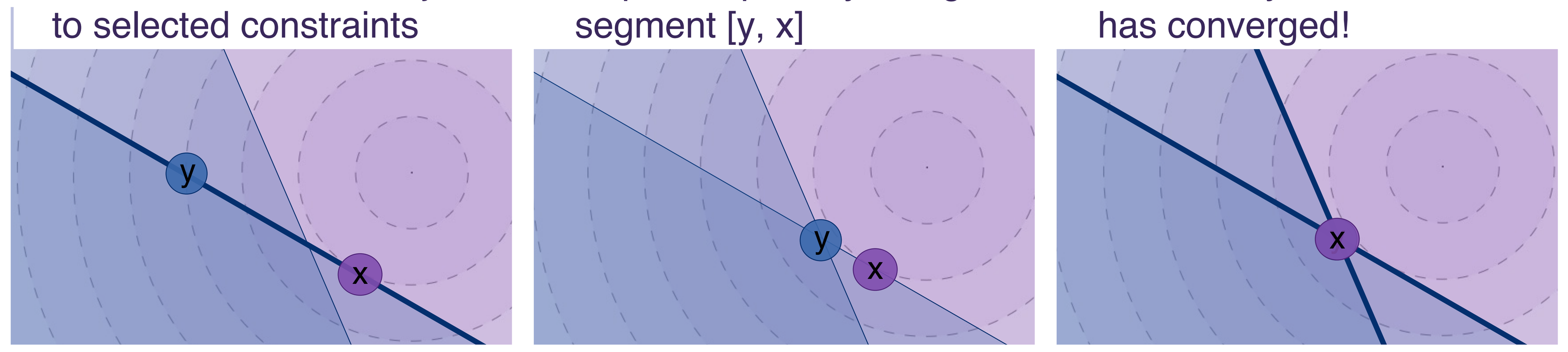
initialize $\mathbf{x} \leftarrow \text{argmin } f(\mathbf{x}')$ **and** $\mathbf{y} \in \mathcal{D}$ ← Feasible region
while not converged **do**
- $\alpha \leftarrow \max\{\alpha' \in [0,1] : \alpha'\mathbf{x} + (1-\alpha')\mathbf{y} \in \mathcal{D}\}$
- $\mathbf{y} \leftarrow \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$
- Choose $\tau \ge 0$ ← Euclidean distance from $\mathbf{y}$ to boundary of h_j
- $\mathcal{C} \leftarrow \{h_j : \text{dist}(h_j, \mathbf{y}) < \tau \vee h_j(\mathbf{x}) = 0\}$
- $\mathbf{x} \leftarrow \text{argmin } f(\mathbf{x}')$ s.t. $h_j(\mathbf{x}') \le 0$ for all $h_j \in \mathcal{C}$

Blitz example with pictures:

1. Initialize y feasible, x minimizer of f
2. Update y as extreme feasible point on [y, x]
3. Select top k constraints with boundaries closest to y

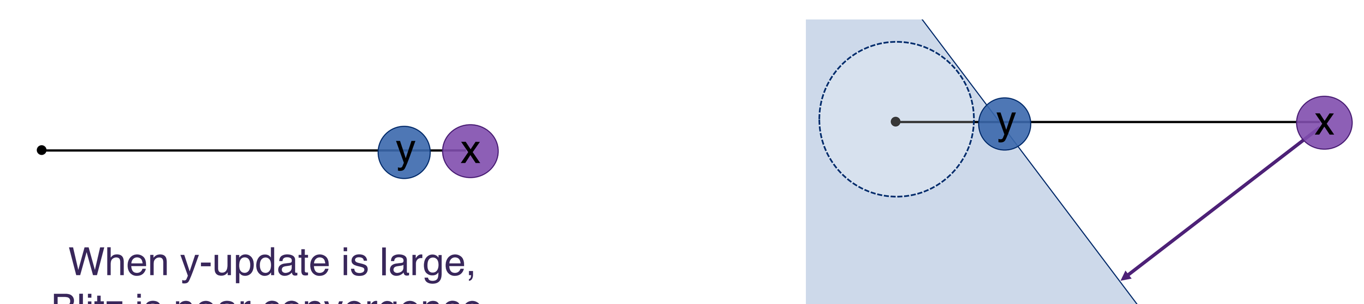


4. Set x to minimize f subject to selected constraints
5. Repeat. Update y along segment [y, x]
5. When x = y, Blitz has converged!



Intuition for Blitz

Intuition for Blitz can be summarized by its y-update:



When y-update is large, Blitz is near convergence.

When y-update is small, geometry of violated constraint ensures large progress in x.

Theoretical Contributions

For iteration t, define suboptimality gap

$$\Delta_t = f(\mathbf{y}_t) - f(\mathbf{x}_t) \ge f(\mathbf{y}_t) - f(\mathbf{x}^\star)\,.$$

Also define:

- f's strong convexity parameter γ
- Radius τ_t (larger τ_t implies larger active set)

Theorem 2.1: Progress at Iteration t

If Algorithm 1 does not converge at iteration $t+1$, then

$$\Delta_{t+1} \le \Delta_t - (\tfrac{\gamma}{2}\tau_t^2\Delta_t^2)^{1/3}\,.$$

Corollary 2.2: Active Set Size for Linear Convergence

For $t \ge 1$, define

$$\Delta_t' = f(\mathbf{y}_t) - f(\mathbf{x}_{t-1})\,,$$

and for some $r \in [0,1)$, suppose we run Blitz choosing radius τ_t as

$$\tau_t = \sqrt{\tfrac{2}{\gamma}(1-r)^3\Delta_t'}\,.$$

The for all $t \ge 1$, we have

$$f(\mathbf{y}_t) - f(\mathbf{x}^\star) \le r^{t-1}\Delta_0\,.$$

Corollary 2.3: Constraint Screening

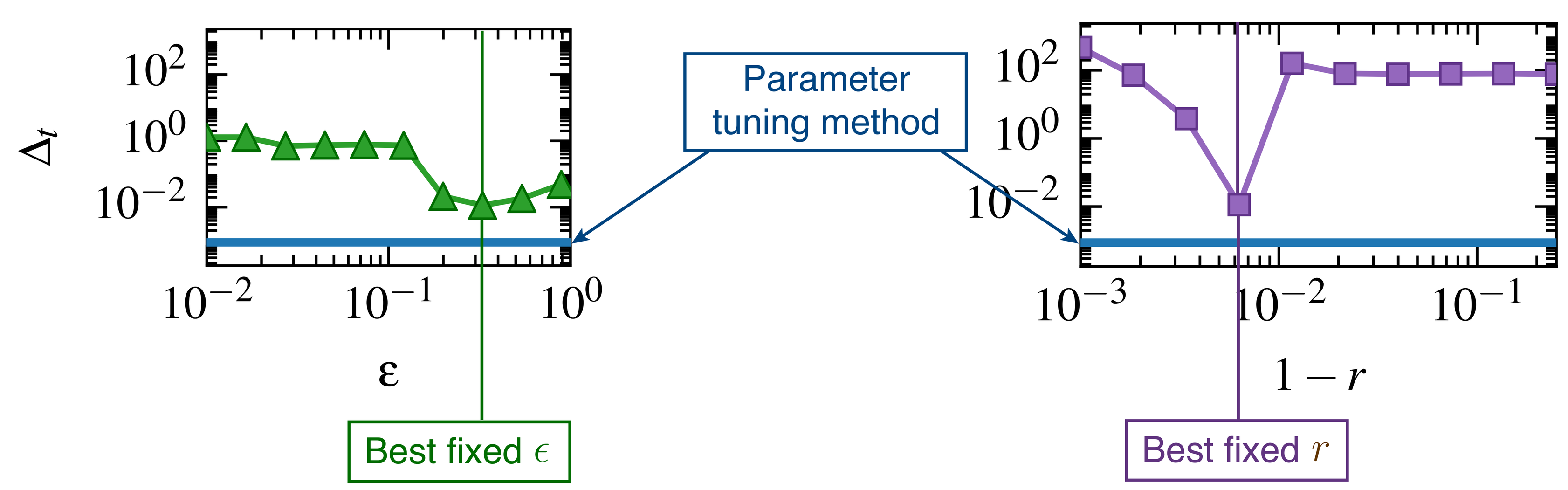
For $t \ge 1$, if

$$\text{dist}(h_j, \mathbf{y}_t) > \sqrt{\tfrac{2}{\gamma}\Delta_t'}\,,$$

then $h_j(\mathbf{x}^\star) < 0$, meaning h_j is guaranteed to be irrelevant to the solution and may be discarded.

Tuning Algorithmic Parameters

Theorem 2.1 can also be used to guide choice of algorithmic parameters ϵ (subproblem stopping tolerance) and r (active set size).



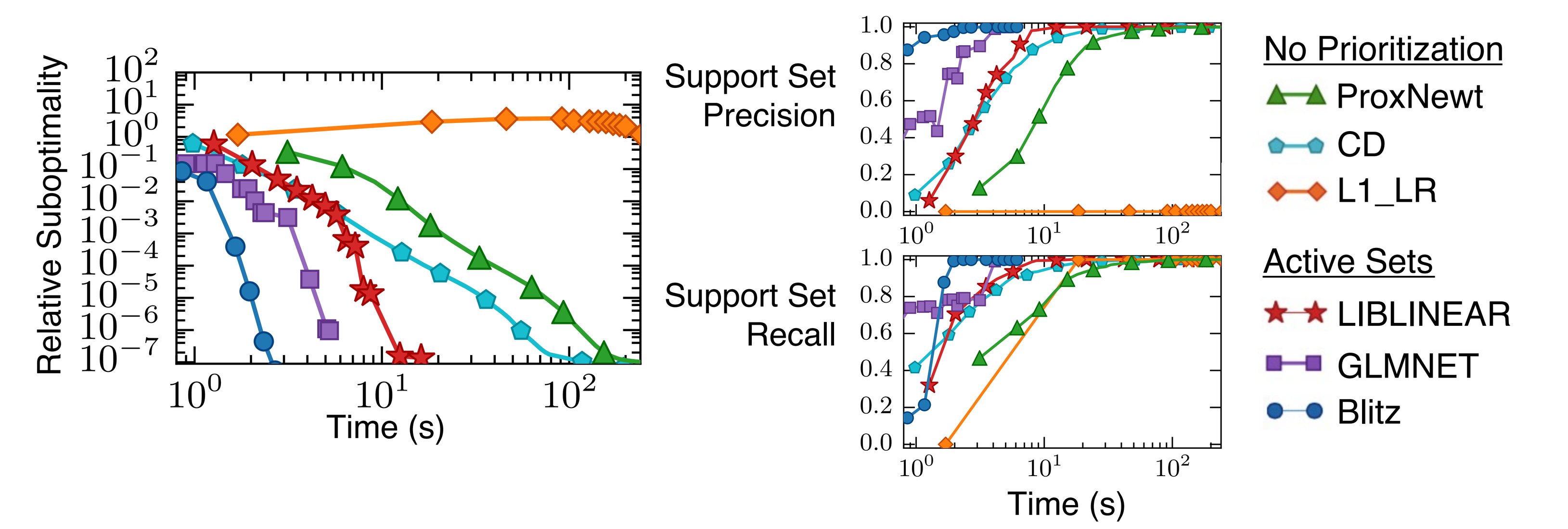
Empirical Results

We apply Blitz to ℓ_1-regularized loss minimization (using duality):

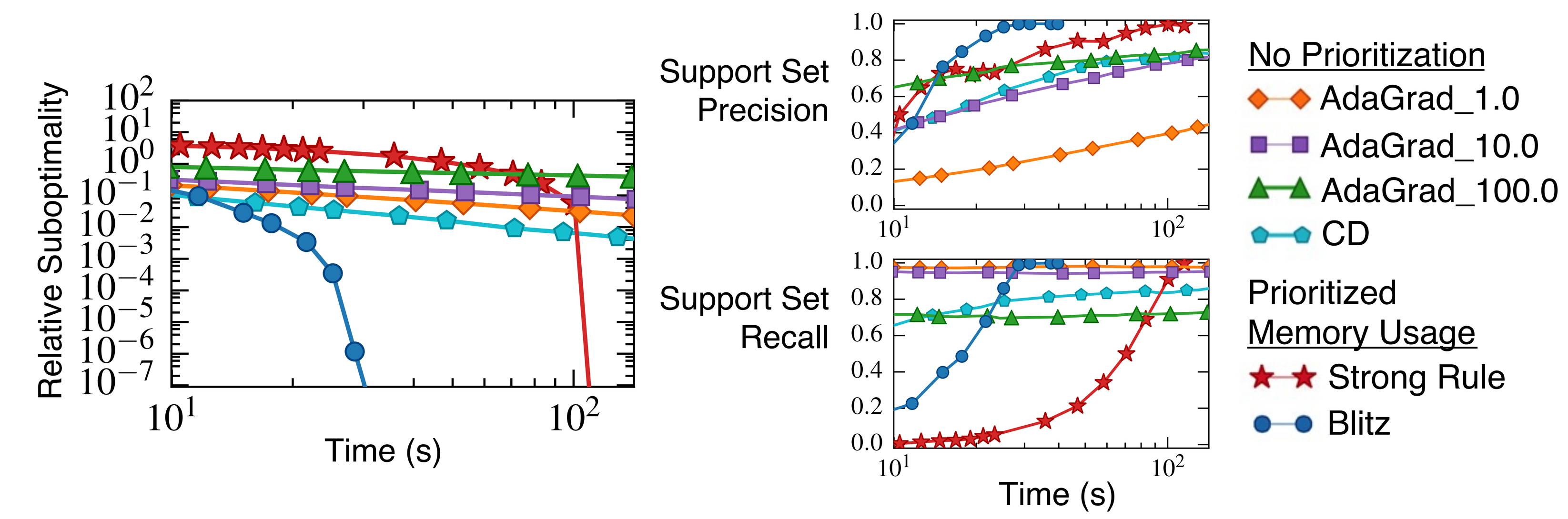
$$\text{minimize} \sum_{i=1}^n \phi_i(\mathbf{a}_i^T\mathbf{w}) + \lambda\|\mathbf{w}\|_1 \longleftrightarrow \begin{aligned} \text{maximize} \quad & \textstyle\sum_{i=1}^n -\phi_i^*(x_i) \\ \text{s.t.} \quad & |\mathbf{A}_j^T\mathbf{x}| \le \lambda \quad j = 1, \ldots, m \end{aligned}$$

- Optimal $\mathbf{w}$ is sparse.
- Solving dual subject to a subset of constraints corresponds to minimizing primal over subset of variables.

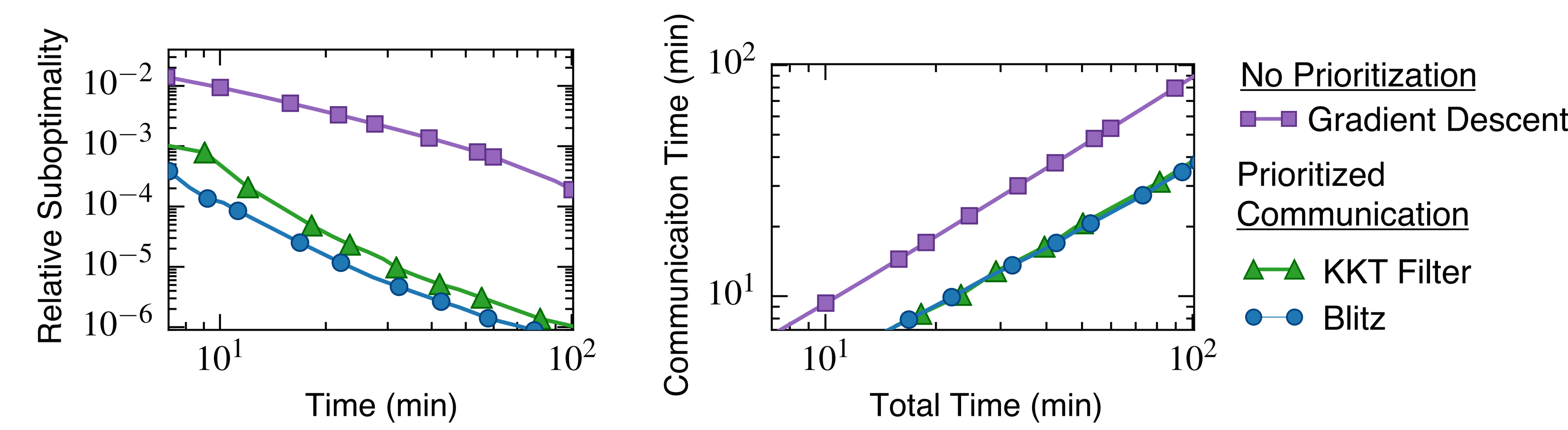
Single machine setting: Experiment with high-dimensional RCV1 dataset.



Limited memory setting: Experiment with 12 GB Webspam dataset and 1 GB memory. Each iteration uses one pass over data to load active set.



Distributed setting: Experiment uses Criteo CTR dataset. Solvers use bulk synchronous gradient descent. Active sets prioritize communication.



Conclusions

We have introduced Blitz, an active set algorithm that

- Selects theoretically justified subproblems to maximize guaranteed progress toward convergence.
- Uses analysis to tune parameters and discard irrelevant constraints.
- Achieves very fast convergence times for ℓ_1-regularized learning.
- Provides a novel proof path for analyzing active set methods.