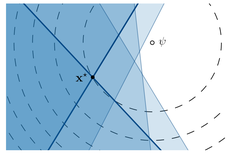
Unified Methods for Exploiting Piecewise Linear Structure in Convex Optimization

Tyler B. Johnson and Carlos Guestrin
University of Washington, Seattle

Introduction

Many convex problems exhibit useful structure at their solutions.

- **SVMs:** Optimal model uninfluenced by non-support vectors.
- **Sparse regression:** Optimal model uses small number of features.
- **Constrained optimization:** Only active constraints determine solution.



Goal

Principled methods that exploit "structure" to achieve fast convergence times.

Existing Approaches

- **Screening:** Identify components (training examples, features) guaranteed to be irrelevant to solution.
- **Working set algorithm:** Solve sequence of subproblems using small subsets of components until convergence.

Main Contributions

1. Formalization of "structure" using piecewise terms in objective function.
2. Fast, principled, and versatile working set algorithm.
3. State-of-the-art screening test, which results from working set analysis.

Piecewise Problem Framework

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\; f(\mathbf{x}) := \psi(\mathbf{x}) + \sum_{i=1}^{m} \phi_i(\mathbf{x})$$

- Each ϕ_i is piecewise; there exists a function $\pi_i : \mathbb{R}^n \to \{1, \dots, p_i\}$, convex functions $\phi_i^1, \dots, \phi_i^{p_i}$ such that for all $\mathbf{x}$,

$$\phi_i(\mathbf{x}) = \phi_i^{\pi_i(\mathbf{x})}(\mathbf{x}).$$

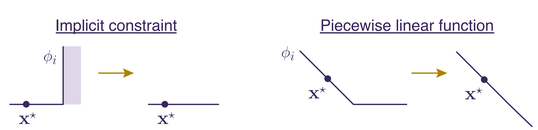
- Subdomain k of ϕ_i is denoted

$$\mathcal{X}_i^k = \{\mathbf{x} : \pi_i(\mathbf{x}) = k\}.$$

- ψ assumed to be γ-strongly convex.

Basic Idea

Selectively replace ϕ_i functions with linear ϕ_i^k such that $\mathbf{x}^\star$ is unchanged.



Proposition 2.1: Exploiting piecewise structure

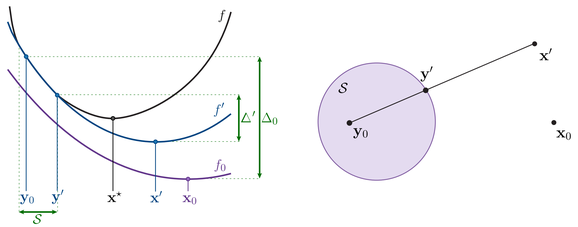
Let $\mathbf{x}^\star$ minimize f. For each $i \in [m]$, define

$$\phi_i^\star = \begin{cases} \phi_i^{\pi_i(\mathbf{x}^\star)} & \text{if } \mathbf{x}^\star \in \operatorname{int}(\mathcal{X}_i^{\pi_i(\mathbf{x}^\star)}), \\ \phi_i & \text{otherwise.} \end{cases}$$

Then $\mathbf{x}^\star$ is also the solution to

$$\underset{\mathbf{x}}{\text{minimize}}\; \psi(\mathbf{x}) + \sum_{i=1}^{m} \phi_i^\star(\mathbf{x}).$$

General Theoretical Result



- Progress toward minimizer of f by minimizing a simpler function f'.
- f' lower bounds f, and $f'(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x}$ in a set $\mathcal{S}$.
- f' upper bounds f_0, which is minimized by $\mathbf{x}_0$.
- For initial $\mathbf{y}_0$, $\Delta_0 := f(\mathbf{y}_0) - f_0(\mathbf{x}_0)$ is suboptimality gap.
- $\mathbf{x}'$ is the minimizer of f'; $\mathbf{y}' := \theta'\mathbf{x}' + (1-\theta')\mathbf{y}_0$ computed via backtracking.
- Result bounds gap $\Delta' := f(\mathbf{y}') - f'(\mathbf{x}')$ in terms of $\mathcal{S}$ and Δ_0.

Lemma 3.1: Guaranteed suboptimality gap progress

$$\Delta' \le (1-\theta')\left[\Delta_0 - \frac{1+\theta'\gamma}{\theta'^2}\min_{\mathbf{z}\notin\operatorname{int}(\mathcal{S})}\left\|\mathbf{z} - \frac{\theta'\mathbf{x}_0+\mathbf{y}_0}{1+\theta'}\right\|^2 - \frac{\theta'}{1+\theta'}\frac{\gamma}{2}\|\mathbf{x}_0-\mathbf{y}_0\|^2\right].$$

Result forms basis for working set algorithm and screening test.

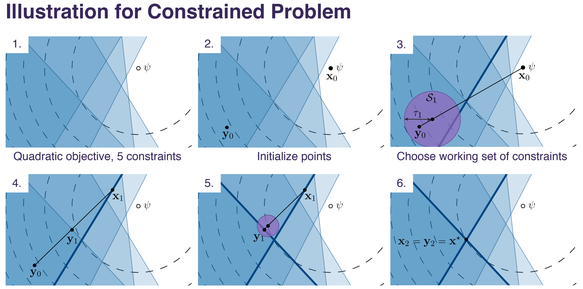
PW-Blitz Working Set Algorithm

Algorithm 1 PW-Blitz

initialize $\mathbf{y}_0$
initialize $\mathbf{x}_0 \leftarrow \operatorname{argmin} f_0'(\mathbf{x})$, where f_0' is a simple lower bound on f
for $t = 1, 2, \dots, T$ **until** converged **do**
  Select $\tau_t > 0, \beta_t \in [0, \frac{1}{2}]$ — Form Subproblem
  $\mathcal{S}_t \leftarrow \operatorname{Ball}(\beta_t\mathbf{x}_{t-1} + (1-\beta_t)\mathbf{y}_{t-1}; \tau_t)$ — Form Subproblem
  for $i = 1, \dots, m$ **do** — Form Subproblem
    if (C1 and C2 and C3) **then** $\phi_{i,t}' := \phi_i^{\pi_i(\mathbf{y}_{t-1})}$ **else** $\phi_{i,t}' := \phi_i$ — Form Subproblem
  end for — Form Subproblem
  $\mathbf{x}_t \leftarrow \operatorname{argmin} f_t'(\mathbf{x}) := \psi(\mathbf{x}) + \sum_{i=1}^m \phi_{i,t}'(\mathbf{x})$ — Solve Subproblem
  $\mathbf{y}_t \leftarrow \underset{\mathbf{x}\in[\mathbf{y}_{t-1},\mathbf{x}_t]}{\operatorname{argmin}} f(\mathbf{x})$ — Backtrack
end for
return $\mathbf{y}_T$

C1. $\phi_i^{\pi_i(\mathbf{y}_{t-1})}(\mathbf{x}) = \phi_i(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}_t$
C2. $\phi_i^{\pi_i(\mathbf{y}_{t-1})}$ lower bounds ϕ_i
C3. $\phi_i^{\pi_i(\mathbf{y}_{t-1})}$ upper bounds $\phi_{i,t-1}'$ in a neighborhood of $\mathbf{x}_{t-1}$

Illustration for Constrained Problem



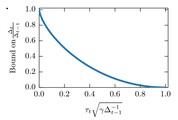
Convergence Analysis

Theorem 3.2: Convergence test with optimal β_t

Let Δ_{t-1} and Δ_t be the suboptimality gaps after iterations $t-1$ and t of PW-Blitz. Assume $\beta_t = \theta_t(1+\theta_t)^{-1}$. Then

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

Links subproblem size (τ_t) to progress toward convergence ($\Delta_{t-1} - \Delta_t$).



Screening Test

Theorem 3.4: Piecewise screening test

Consider any $\mathbf{x}_0$, $\mathbf{y}_0$ such that $\mathbf{x}_0$ minimizes a γ-strongly convex function f_0 that lower bounds f. Define $\Delta_0 = f(\mathbf{y}_0) - f_0(\mathbf{x}_0)$. For any $i \in [m]$, let $k = \pi_i(\mathbf{y}_0)$. If

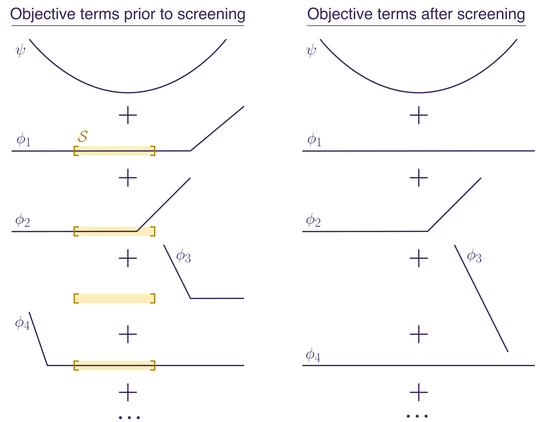
$$\mathcal{S} := \left\{\mathbf{x} : \left\|\mathbf{x} - \tfrac{1}{2}(\mathbf{x}_0+\mathbf{y}_0)\right\| \le \sqrt{\tfrac{1}{\gamma}\Delta_0 - \tfrac{1}{4}\|\mathbf{x}_0-\mathbf{y}_0\|^2}\right\} \subseteq \operatorname{int}(\mathcal{X}_i^k),$$

then $\mathbf{x}^\star \in \operatorname{int}(\mathcal{X}_i^k)$. ϕ_i may be replaced with ϕ_i^k in (P) without affecting $\mathbf{x}^\star$.

Proof applies Lemma 3.1 to guarantee $\mathbf{x}^\star \in \mathcal{S}$.

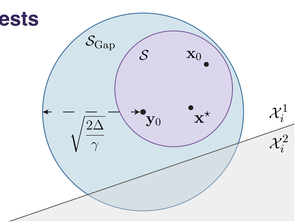
Visual Example: Support Vector Machines

$$f(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|^2 + C\sum_{i=1}^m (1 - b_i\langle\mathbf{a}_i, \mathbf{x}\rangle)_+$$



Relation to Prior Screening Tests

- Prior tests apply to small classes of objectives.
- Prior tests may fail to screen components that Theorem 3.4 screens successfully.
- Many prior tests only apply as a preprocessing step.



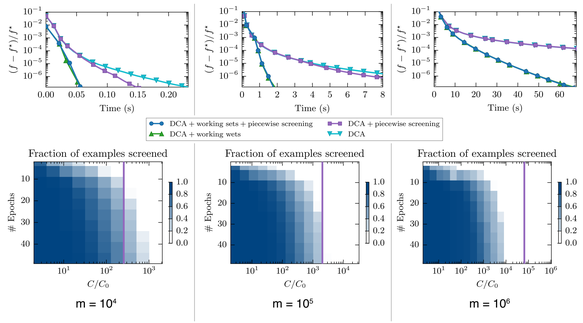
Empirical Results

Compared scalability of screening and working sets as m increases.

- **DCA:** Dual stochastic coordinate ascent.
- **DCA + working sets:** PW-Blitz; DCA solves each subproblem.
- **DCA + screening:** DCA with Theorem 3.4 applied after every 5 epochs.
- **DCA + working sets + screening:** PW-Blitz with screening.

Support Vector Machines

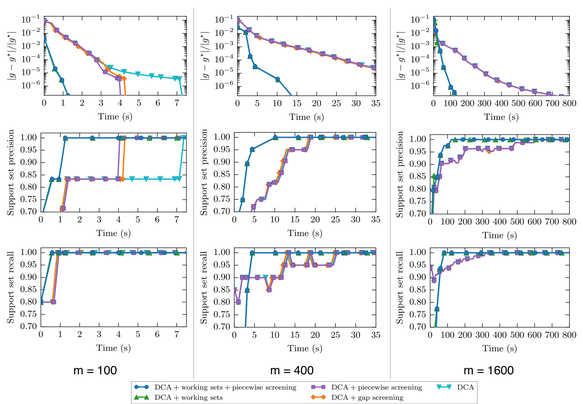
Comparisons use Higgs boson dataset (8010 features).



Group Lasso

$$g(\boldsymbol{\omega}) = \tfrac{1}{2}\left\|\mathbf{A}\boldsymbol{\omega} - \mathbf{b}\right\|^2 + \lambda\sum_{i=1}^m \left\|\boldsymbol{\omega}_{\mathcal{G}_i}\right\|_2$$

Tests use Allstate insurance claim dataset (250k examples, 29k features).



Conclusions

- We developed a fast, principled working set algorithm and state-of-the-art screening test using unified theory.
- Empirical results demonstrate working set algorithms are more effective than screening, especially as problem size m increases.