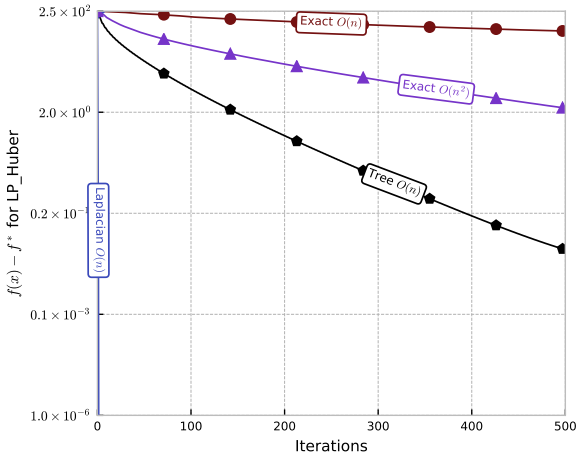
Newton-Laplace Updates for Block Coordinate Descent

Machine learning tasks are often formulated as minimizing an objective function f with input $x \in \mathbb{R}^n$. When n is large, block coordinate descent methods are usually efficient as they only update a subset of the variables at each iteration. In addition to using the gradient information to compute the update, several authors have explored incorporating second-order information to obtain faster convergence [5, 2]. Specifically, the blockwise-Newton update at iteration k has the form $x_b^{k+1} = x_b^k - \alpha^k \left[\nabla^2_{bb} f(x^k)\right]^{-1} \nabla_b f(x^k)$ for a selected block $b \subseteq [n] = \{1, \ldots, n\}$ with a step size α^k. Since this method requires solving a linear system of dimension $|b|$, it incurs an iteration cost of $O(|b|^3)$ using generic matrix factorization methods. This can prohibit the use of large blocks that can potentially lead to faster convergence. Nutini *et al.*[4] show that when the chosen block's sparsity pattern has a tree structure, "message-passing" algorithms can be used to solve the system in linear time. Srinivasan and Todorov [7] exploit the width of the Hessian's computation graph to speed up the Newton update. However, there could still be a limit to how large the blocks can grow until these structural constraints are violated.

In this work, we consider an alternative structural assumption on the blocks that allows the use of fast solvers for blockwise-Newton updates. For an undirected weighted graph $\mathcal{G}$, we denote its adjacency matrix by W and the Laplacian matrix by L. Whenever the (sub-)Hessian of f is a Laplacian matrix, we can leverage near-linear time Laplacian solvers [6, 3] that perform approximate sub-sampled Cholesky factorization to circumvent the cubic iteration cost. An application would be the classic graph-based semi-supervised learning problem: given a labeling $x = (x_l, x_u)$, the label propagation algorithm [9] minimizes the objective $f(x_u) = \sum_{i \in [n], j \in u} w_{ij} h(x_i - x_j) + \frac{1}{2} \sum_{i,j \in u} w_{ij} h(x_i - x_j)$, where $u \subseteq [n]$ is the set of unlabeled examples and $l = [n] \setminus u$. When $h(z) = z^2$, the analytical solution can be obtained from $x_u = L_{uu}^{-1}(W_{ul} x_l)$ [1] where L_{uu} is the Laplacian on the graph of the unlabeled examples. The Laplacian solvers allow us to use arbitrarily large block sizes while maintaining an iteration cost in only $\tilde{O}(|b|)$. We also show that when replacing $h(\cdot)$ in f with the Huber loss, the (sub-)Hessian of f still has a Laplacian-like structure, and thus again allows us to use blockwise-Newton updates to obtain fast convergence with cheap iterations.

In the figure below we show experimental results of L2-regularized label propagation on the "two moons" dataset [8] with pairwise distance measured in the Huber loss. The Newton updates using exact solvers use fixed-size blocks that are chosen such that the iteration costs are $O(n)$ and $O(n^2)$, respectively, where $n = 1900$ is the number of variables. Although we are able to obtain a slightly faster convergence, our iteration cost has increased significantly. If we instead select the blocks using tree partitioning[4], we are able to use much larger blocks to improve the convergence with $O(n)$ iteration cost. Finally, the Laplacian solver allows us to use full blocks with $|b| = n$ and converges in only a few steps without sacrificing iteration time complexity.



References

[1] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 11 label propagation and quadratic criterion. 2006.

[2] Kimon Fountoulakis and Rachael Tappenden. A flexible coordinate descent method. *arXiv preprint arXiv:1507.03713*, 2015.

[3] Rasmus Kyng and Sushant Sachdeva. Approximate gaussian elimination for laplacians-fast, sparse, and simple. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2016.

[4] Julie Nutini, Issam Laradji, and Mark Schmidt. Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *arXiv preprint arXiv:1712.08859*, 2017.

[5] Zheng Qu, Peter Richtárik, Martin Takác, and Olivier Fercoq. Sdna: stochastic dual newton ascent for empirical risk minimization. In *International Conference on Machine Learning*, 2016.

[6] Daniel A Spielman et al. Laplacians. jl, 2017.

[7] Akshay Srinivasan and Emanuel Todorov. Graphical newton. *arXiv preprint arXiv:1508.00952*, 2015.

[8] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, 2004.

[9] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.