Neon: Nuclear Norm to Beat Muon

Alexey Kravatskiy, Ivan Kozyrev, Nikolay Kozlov, Alexander Vinogradov
Optimization Class Project. MIPT

Introduction

Recent advances in optimization techniques have highlighted the benefits of leveraging the matrix structure of neural network weights during training. Optimizers like Muon (Jordan et al.) and Shampoo (Gupta, Koren and Singer) have shown promise, but at a significantly higher computational cost than traditional methods like Adam. To bridge this gap, we propose Neon, a new optimizer that builds upon the framework of Bernstein and Newhouse. By using alternative norms, such as kernel norm (Kernel-Neon) or custom $F*$ norm ($F*$-Neon), we induce low-rank update matrices that enable more efficient computation. We evaluate the performance of Neon, Muon, and Adam on multilayer perceptrons, CIFAR10, and NanoGPT, and demonstrate [resting results here].

Neon's update rule

Bernstein and Newhouse suggest obtaining the update step for a weight matrix W as a solution to the optimization problem:

$$\langle G, \delta W\rangle + \frac{\lambda}{2}\|\delta W\|^2 \to \min_{\delta W}, \tag{1}$$

where G is a gradient-like matrix obtained via backpropagation. Setting norm to RMS-to-RMS norm (scaled version of Spectral norm) produces Muon. We consider two different choices instead:

1. Choosing kernel norm, $\|\cdot\|_*$, produces rank-1 update, defined by

$$\delta W = -\frac{1}{\lambda} u_1 \sigma_1 v_1^T, \tag{2}$$

   where σ_1 is largest singular value of G, and u_1, v_1 are corresponding singular values.

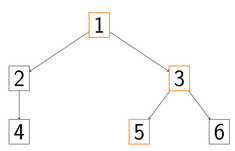
2. Choosing $F*$ norm, defined by $\|\cdot\|_{F*} = (\|\cdot\|_* + \|\cdot\|_F)/2$, produces a relatively small-rank update, defined by

$$\delta W = -\frac{1}{\lambda} U D V^T \tag{3}$$

   with $D = \mathbf{diag}(d_i)$, where $d_i = [\sigma_i - \tau]_+$ and τ is given by

$$\sum_{i=1}^{n} [\sigma_i - \tau]_+ = \tau\,.$$

Hierarchical selection



- $\mathcal{G} = \{\{4\}, \{5\}, \{6\}, \{2,4\}, \{3,5,6\}, \{1,2,3,4,5,6\}\}$
- nonzero variables form a rooted and connected subtree
  - if node is selected, so are its ancestors
  - if node is not selected, neither are its descendants

Algorithm

We solve this problem using an ADMM lasso implementation:

```
prox_f = @(v,rho) (rho/(1 + rho))*(v - b) + b;
prox_g = @(v,rho) (max(0, v - 1/rho) - max(0, -v - 1/rho));

AA = A*A';
L  = chol(eye(m) + AA);

for iter = 1:MAX_ITER
    xx = prox_g(xz - xt, rho);
    yx = prox_f(yz - yt, rho);

    yz = L \ (L' \ (A*(xx + xt) + AA*(yx + yt)));
    xz = xx + xt + A'*(yx + yt - yz);

    xt = xt + xx - xz;
    yt = yt + yx - yz;
end
```

Line search

If L is not known (usually the case), can use the following line search:

---

given x^k, λ^{k-1}, and parameter $\beta \in (0,1)$.
Let $\lambda := \lambda^{k-1}$.
repeat
1. Let $z := \mathbf{prox}_{\lambda g}(x^k - \lambda \nabla f(x^k))$.
2. **break if** $f(z) \le \hat{f}_\lambda(z, x^k)$.
3. Update $\lambda := \beta\lambda$.

return $\lambda^k := \lambda$, $x^{k+1} := z$.

---

typical value of β is $1/2$, and

$$\hat{f}_\lambda(x,y) = f(y) + \nabla f(y)^T(x-y) + (1/2\lambda)\|x-y\|_2^2$$

Convergence proof

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Numerical example

Consider a numerical example with $f(x) = \|Ax - b\|_2^2$ with $A \in \mathbf{R}^{10\times 100}$ and $b \in \mathbf{R}^{10}$. Entries of A and b are generated as independent samples from a standard normal distribution. Here, we have chosen λ using cross validation.

Results

On this numerical example, the ADMM method converges quickly. We give two realizations corresponding to different parameters A and b.

figures/pwl_error_fbest_realize-eps-converted-to.pdf

Conclusion

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Acknowledgements

This material is based upon work supported by the X Fellowship and my mom.