Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 23: Adaptive Gradient Methods

Stochastic Optimization in Machine Learning

In machine learning, we observe training data (x_i, y_i) for $i = 1, \ldots, n$, and would like to learn a model parameter w of the form

$$\min_{w \in C} \left[\frac{1}{n} \sum_{i=1}^{n} f_i(w) + g(w) \right].$$

More generally, we can write this optimization problem as:

$$\min_{w \in C} \phi(w), \quad \phi(w) = f(w) + g(w), \qquad f(w) = \mathbf{E}_{\xi \sim D} f(\xi, w), \qquad (1)$$

where ξ is a random variable, drawn from a distribution D.

Coordinate-wise Learning Rate

Given a minibatch B, and positive definite matrix Λ, we define

$$f_B(w) = \frac{1}{|B|} \sum_{\xi \in B} f(\xi, w),$$

and

$$\mathrm{prox}_{\Lambda,g}(w) = \arg\min_z \left[\frac{1}{2}(z - w)^\top \Lambda^{-1}(z - w) + g(z) \right].$$

Consider the following general proximal stochastic gradient method below with $B_t \sim D$:

$$w^{(t)} = \mathrm{prox}_{\Lambda_t,g}(w^{(t-1)} - \Lambda_t \nabla f_{B_t}(w^{(t-1)}),$$

where we replace the coordinate independent learning rate η_t by a diagonal matrix $\Lambda_t = \mathrm{diag}(\eta_{t,1}, \ldots, \eta_{t,d})$. Here $\eta_{t,j}$ is the learning rate for the j-th coordinate at time t.

General Property

If we define

$$Q_{\Lambda,B}(w;w') = f(w') + \nabla f_B(w')^\top (w - w') + \frac{1}{2}\|w - w'\|^2_{\Lambda^{-1}} + g(w),$$

where

$$\|u\|_\Lambda^2 = u^\top \Lambda u.$$

Proposition

Assume that $f(w)$ and $g(w)$ are convex. Then

$$f(w') + g(w) \le Q_{\Lambda,B}(w;w') + \frac{1}{2}\|\nabla f_B(w')\|_\Lambda^2.$$

Proof

The first inequality follows from the same argument of Lecture 19. For the second inequality, we note that

$$\begin{aligned} f(w') + g(w) \leq & f(w') + g(w) + \frac{1}{2}\|w - w' + \Lambda\nabla f_B(w')\|_{\Lambda^{-1}}^2 \\ = & Q_{\Lambda,B}(w; w) + \frac{1}{2}\|\nabla f_B(w')\|_{\Lambda}^2. \end{aligned}$$

This proves the result.

Convergence Result

Theorem

Let $\Delta_j = \max_{w,w' \in C} |w - w'|_j$, *and* $\Delta = \mathrm{diag}(\Delta_1, \ldots, \Delta_d)$. *If we take* $\Lambda_1 \geq \Lambda_2 \geq \cdots$, *then*

$$\mathbf{E} \sum_{t=1}^{T} \phi(w^{(t-1)}) \leq T\phi(w) + \mathbf{E}[g(w^{(0)}) - g(w^{(T)})] + \frac{1}{2}\mathbf{E}\ \mathit{trace}(\Delta^2 \Lambda_T^{-1}) + \frac{1}{2} \sum_{t=1}^{T}$$

Proof

We have from Proposition 1

$$\begin{aligned}
&f(w^{(t-1)}) + g(w^{(t)}) \leq Q_{\Lambda_t,B_t}(w^{(t)}; w^{(t-1)}) + \frac{1}{2}\|\nabla f_{B_t}(w^{(t-1)})\|_{\Lambda_t}^2 \\
\leq &Q_{\Lambda_t,B_t}(w; w^{(t-1)}) - \frac{1}{2}(w - w^{(t)})\Lambda_t^{-1}(w - w^{(t)}) + \frac{1}{2}\|\nabla f_{B_t}(w^{(t-1)})\|_{\Lambda_t}^2 \\
\leq &f(w^{(t-1)}) + \nabla f_{B_t}(w^{(t-1)})^\top (w - w^{(t-1)}) + g(w) \\
&+ \frac{1}{2}\|w - w^{(t-1)}\|_{\Lambda_t^{-1}}^2 - \frac{1}{2}\|w - w^{(t)}\|_{\Lambda_t^{-1}}^2 + \frac{1}{2}\|\nabla f_{B_t}(w^{(t-1)})\|_{\Lambda_t}^2.
\end{aligned}$$

We obtain

$$\mathbf{E}_{B_t}[f(w^{(t-1)}) + g(w^{(t)})] \leq \phi(w) + \frac{1}{2}\mathbf{E}_{B_t}\left[\|w - w^{(t-1)}\|_{\Lambda_t^{-1}}^2 - \|w - w^{(t)}\|_{\Lambda_t^{-1}}^2\right]$$

Proof: continues

By summing over t, and noticing that (we take $\Lambda_0^{-1} = 0$):

$$\begin{aligned}
&\sum_{t=1}^{T} \left[\|w - w^{(t-1)}\|_{\Lambda_t^{-1}}^2 - \|w - w^{(t)}\|_{\Lambda_t^{-1}}^2 \right] \\
\leq &\sum_{t=1}^{T} \|w - w^{(t-1)}\|_{\Lambda_t^{-1} - \Lambda_{t-1}^{-1}}^2 \\
\leq &\sum_{t=1}^{T} \mathrm{trace}(\Delta^2 (\Lambda_t^{-1} - \Lambda_{t-1}^{-1})) = \mathrm{trace}(\Delta^2 \Lambda_T^{-1}),
\end{aligned}$$

we obtain the bound.

AdaGrad Algorithm

Algorithm 1: AdaGrad

```
Input: $\phi(\cdot)$, learning rates $\eta$ , $\epsilon > 0$, $w^{(0)}$
Output: $w^{(T)}$
$\tilde{g}_0^2 = [\epsilon, \ldots, \epsilon]$
for $t = 1, 2, \ldots, T$ do
    Randomly pick $B_t \sim D$
    Let $g_t = \nabla_w f_{B_t}(w^{(t-1)})$
    Let $\tilde{g}_t^2 = \tilde{g}_{t-1}^2 + g_t^2$
    Let $\Lambda_t = \eta \mathrm{diag}(\tilde{g}_t^{-1})$
    Let $w^{(t)} = \mathrm{prox}_{\Lambda_t g}(w^{(t-1)} - \Lambda_t g_t)$
Return: $w^{(T)}$
```

Theory

Corollary

If for some $\eta > 0$, *we take* $\eta_{t,j}$ *in* Λ_t *for* $j = 1, \ldots, d$ *as*

$$\eta_{t,j} = \frac{\eta \Delta_j}{\sqrt{\epsilon + \sum_{s=1}^{t} [\nabla f_{B_s}(w^{(s-1)})]_j^2}},$$

then

$$\mathbf{E} \sum_{t=1}^{T} \phi(w^{(t-1)}) \leq T\phi(w) + \mathbf{E}[g(w^{(0)}) - g(w^{(T)})]$$

$$+ (0.5\eta^{-1} + \eta)\mathbf{E} \sum_{j=1}^{d} \Delta_j \sqrt{\epsilon + \sum_{s=1}^{T} [\nabla f_{B_s}(w^{(s-1)})]_j^2}.$$

Proof

We have

$$\mathrm{trace}(\Delta^2\Lambda_T^{-1}) = \eta^{-1}\sum_{j=1}^{d}\Delta_j\sqrt{\epsilon+\sum_{s=1}^{T}[\nabla f_{B_s}(w^{(s-1)})]_j^2},$$

and

$$\sum_{t=1}^{T}\|\nabla f_{B_t}(w^{(t-1)})\|_{\Lambda_t}^2 = \eta^2\sum_{t=1}^{T}\sum_{j=1}^{d}\Delta_j^2\frac{\eta_{t,j}^{-2}-\eta_{t-1,j}^{-2}}{\eta_{t,j}^{-1}}$$

$$\leq 2\eta^2\sum_{t=1}^{T}\sum_{j=1}^{d}\Delta_j^2\frac{\eta_{t,j}^{-2}-\eta_{t-1,j}^{-2}}{\eta_{t,j}^{-1}+\eta_{t-1,j}^{-1}} = 2\eta^2\sum_{t=1}^{T}\sum_{j=1}^{d}\Delta_j^2(\eta_{t,j}^{-1}-\eta_{t-1,j}^{-1})$$

$$=2\eta\,\mathrm{trace}(\Delta^2\Lambda_T^{-1}).$$

This implies the bound.

RDA version

Algorithm 2: AdaGrad-RDA

```
Input: $f(\cdot)$, $g(\cdot)$, $w^{(0)}$, $\eta_0, \eta_1, \eta_2, \ldots$
    $h(w)$ (default is $h(w) = 0.5\eta_0 \|w\|_2^2$)
Output: $w^{(T)}$
Let $\tilde{\alpha}_0 \in \partial h(w^{(0)})$
Let $\tilde{\Lambda}_0 = 0$
$\tilde{g}_0^2 = [\epsilon, \ldots, \epsilon]$
for $t = 1, 2, \ldots, T$ do
    Randomly select a minibatch $B_t$ of $m$ independent samples from $D$
    Let $g_t = \nabla_w f_{B_t}(w^{(t-1)})$
    Let $\tilde{g}_t^2 = \tilde{g}_{t-1}^2 + g_t^2$
    Let $\Lambda_t = \eta \mathrm{diag}(\tilde{g}_t^{-1})$
    Let $\tilde{\alpha}_t = \tilde{\alpha}_{t-1} - \Lambda_t g_t$
    Let $\tilde{\Lambda}_t = \tilde{\Lambda}_{t-1} + \Lambda_t$
    Let $w^{(t)} = \mathrm{prox}_{\tilde{\Lambda}_t g}(\tilde{\alpha}_t)$
Return: $w^{(T)}$
```

RMSprop

Algorithm 3: RMSprop

```
Input: $\phi(\cdot)$, learning rates $\eta$ , $\rho$ (default is 0.9),$\epsilon > 0$, $w^{(0)}$
Output: $w^{(T)}$
$\tilde{g}_0^2 = [\epsilon, \ldots, \epsilon]$
for $t = 1, 2, \ldots, T$ do
    Randomly pick $B_t \sim D$
    Let $g_t = \nabla_w f_{B_t}(w^{(t-1)})$
    Let $\tilde{g}_t^2 = \rho \tilde{g}_{t-1}^2 + (1 - \rho) g_t^2$
    Let $\Lambda_t = \eta \mathrm{diag}(\tilde{g}_t^{-1})$
    Let $w^{(t)} = \mathrm{prox}_{\Lambda_t g}(w^{(t-1)} - \Lambda_t g_t)$
Return: $w^{(T)}$
```

Automatic Tuning

```
**Algorithm 4:** AdaDelta
**Input**: $\phi(\cdot)$, learning rates $\eta$, $\rho$, $\epsilon > 0$, $w^{(0)}$
**Output**: $w^{(T)}$
Let $\tilde{g}_0^2 = 0$
Let $\eta_0^2 = 0$
**for** $t = 1, 2, \ldots, T$ **do**
    Randomly pick $B_t \sim D$
    Let $g_t = \nabla_w f_{B_t}(w^{(t-1)})$
    Let $\tilde{g}_t^2 = \rho \tilde{g}_{t-1}^2 + (1 - \rho) g_t^2$
    Let $\Lambda_t = \mathrm{diag}\left(\sqrt{\epsilon + \eta_{t-1}^2} / \sqrt{\epsilon + \tilde{g}_t^2}\right)$
    Let $w^{(t)} = \mathrm{prox}_{\Lambda_t g}(w^{(t-1)} - \Lambda_t g_t)$
    Let $\eta_t^2 = \rho \eta_{t-1}^2 + (1 - \rho)(w^{(t)} - w^{(t-1)})^2$
**Return**: $w^{(T)}$
```

AdaMD

Algorithm 5: AdaMD

```
Input: $f(\cdot)$, $g(\cdot)$, $w^{(0)}$, $\eta_0$, $c$, $p$ (default is $\lceil n/m \rceil$)
Output: $w^{(T)}$
Let $\Lambda_0 = \eta_0 \Lambda$
Let $\tilde{g}^2 = 0$
Let $w_{\min} = w^{(0)}$
Let $w_{\max} = w^{(0)}$
Let $q = 0$
for $t = 1, 2, \ldots, T$ do
    Randomly pick $B_t \sim D$
    Let $g_t = \nabla_w f_{B_t}(w^{(t-1)})$
    Let $\tilde{g}^2 = \tilde{g}^2 + g_t^2$
    Let $\Lambda_t = \Lambda_{t-1}$
    Let $w^{(t)} = \mathrm{prox}_{\Lambda_t g}(w^{(t-1)} - \Lambda_t g_t)$
    Let $w_{\min} = \min(w_{\min}, w^{(t)})$
    Let $w_{\max} = \max(w_{\max}, w^{(t)})$
    Let $q = q + 1$
    if $q >= p$ then
        Let $\Lambda_t = \mathrm{diag}((c(w_{\max} - w_{\min}) + \sqrt{\epsilon})/\sqrt{\epsilon + \tilde{g}^2})$
        Let $\tilde{g}^2 = 0$
        Let $w_{\min} = w^{(t)}$
        Let $w_{\max} = w^{(t)}$
        Let $q = 0$
Return: $w^{(T)}$
```

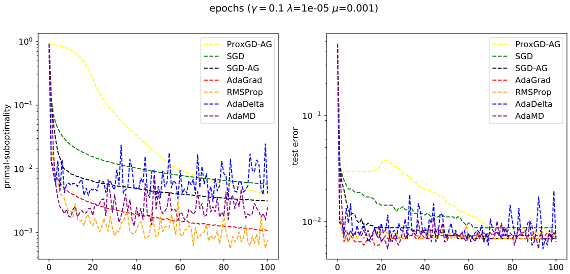
Empirical Studies

We study the smoothed hinge loss function $\phi_\gamma(z)$ with $\gamma = 1$, and solves the following $L_1 - L_2$ regularization problem:

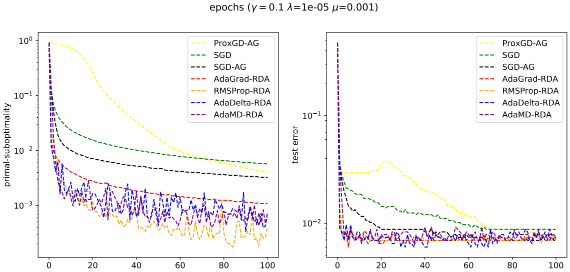
$$\min_w \left[\underbrace{\frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(w^\top x_i y_i)}_{f(w)} + \underbrace{\frac{\lambda}{2} \|w\|_2^2 + \mu \|w\|_1}_{g(w)} \right].$$

We compare different algorithms with constant learning rate.

Comparisons (proximal)



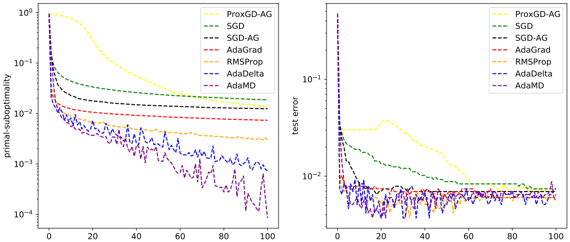
epochs ($\gamma = 0.1$ λ=1e-05 μ=0.001)



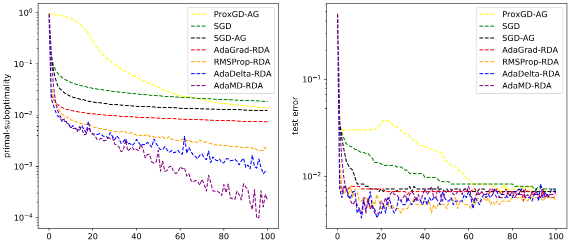
epochs ($\gamma = 0.1$ λ=1e-05 μ=0.001)

Comparisons (no-proximal)

epochs ($\gamma = 0.1$ λ=1e-16 μ=1e-16)



epochs ($\gamma = 0.1$ λ=1e-16 μ=1e-16)



Summary

Coordinate-wise Gradients

- Using gradient to estimate variance
- Only applies to stochastic optimization

AdaGrad

- Search a larger parameter space quickly.
- L_∞-norm rather than L_2-norm
- Might hurt generalization in some cases

Faster Convergence, but less robust