Large-Scale Optimization

L9. PEGASOS
Today:

Pegasos: SGD applied to SVM
Lagrangian Duality

Primal Problem
\[
\min_{x \in \mathbb{R}^n} f(x) \\
\text{s.t. } c_i(x) \geq 0, \ i = 1, 2, \ldots, m.
\]

Lagrangian Function
\[
\mathcal{L}(x; \alpha) := f(x) - \alpha^T c(x) = f(x) - \sum_{i=1}^{m} \alpha_i c_i(x)
\]
\[
\alpha_i \geq 0 \ i = 1, 2, \ldots, m
\]

Dual Problem
\[
\max_{\alpha \in \mathbb{R}^m, \alpha \geq 0} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x; \alpha)
\]

Dual Function \( q(\alpha) \)

- The dual objective \( q \) is concave
- The effective domain of \( q \) is convex
Duality

Weak Duality: for primal feasible $x$ and dual feasible $\alpha$

$$q(\alpha) \leq f(x)$$

Strong Duality: for optimal solutions pair $(x^*, \alpha^*)$ that satisfies KKT conditions

$$q(\alpha^*) = f(x^*)$$

Strong duality holds if one of the following conditions holds:

- $f$ convex, $f^*$ finite, constraints are linear
- $f$ convex, $-c_i$ convex, there exists an interior point (i.e. Slater’s condition hold)
**Strong Convexity**

\[ f(x) = s(x) + c(x) \]

Is \( f \) strongly convex in \( x \) when ... ?

- \( s \) is strongly convex in \( x \)
- \( c \) is convex in \( x \)

\[
\begin{align*}
  s(y) & \geq s(x) + g_s(x)^T (y - x) + \frac{\sigma}{2} \|y - x\|^2, & g_s(x) & \in \partial s(x) \\
  c(y) & \geq c(x) + g_c(x)^T (y - x), & g_c(x) & \in \partial c(x) \\
  s(y) + c(y) & \geq s(x) + c(y) + (g_s(x) + g_c(x))^T (y - x) + \frac{\sigma}{2} \|y - x\|^2,
\end{align*}
\]

E.g.  \( f(w) = \frac{1}{m} \|y - Xw\|_2^2 + \lambda \|w\|_2^2 \)  \( f(w) = \frac{1}{m} \|y - Xw\|_2^2 + \lambda \|w\|_1 \)
Pegasos [Shalev-Shwartz, Singer, Srebro 07]

\[
\min_{w \in \mathbb{R}^n} \frac{\lambda}{2} \| w \|_2^2 + \frac{1}{m} \sum_{i=1}^{m} \ell(w; x_i, y_i)
\]

Consider minimizations of instantaneous approximate objective functions

\[
\min_{w \in \mathbb{R}^n} f(w; A_t) := \frac{\lambda}{2} \| w \|_2^2 + \frac{1}{k} \sum_{(x,y) \in A_t} \ell(w; (x, y))
\]

For us, it is equal to consider

\[
\min_{w \in \mathbb{R}^n} \frac{\lambda}{2} \| w \|_2^2 + \mathbb{E}_{(x,y)}[\ell(w; x, y)]
\]

and stochastic subgradients constructed with k examples
Subgradient of Hinge Loss

\[ \ell(w; (x, y)) = \max\{1 - y\langle w, x \rangle, 0\} \]

\[ g(w; (x, y)) = ? \]

\[ [g(w; (x, y))]_i = \begin{cases} 
-yx_i & \text{if } y\langle x, y \rangle < 1 \\
0 & \text{if } y\langle x, y \rangle > 1 \\
[0, -yx_i] \text{ or } [-yx_i, 0] & \text{o.w.}
\end{cases} \]
Pegasos Algorithm

\[ \min_{w \in \mathbb{R}^n} \frac{\lambda}{2} \| w \|_2^2 + \mathbb{E}_{(x,y)}[\ell(w; x, y)] \]

\[ w_{t+1} = \prod_B (w_t - \eta_t \nabla_t) \]

\[ \eta_t = \frac{1}{\lambda t} \]

\[ B = \{ w : \| w \|_2 \leq 1/\sqrt{\lambda} \} \]

\[ \nabla_t = \lambda w_t - \frac{1}{|A_t|} \sum_{(x,y) \in A_t^+} yx \]

\[ A_t^+ = \{ (x, y) \in A_t : y \langle w, x \rangle < 1 \} \]
Boundedness of $||w||$

How to obtain $B = \{w : ||w||_2 \leq 1/\sqrt{\lambda}\}$?
Theorem 1. Assume that for all \((x, y) \in S\) the norm of \(x\) is at most \(R\). Let \(w^*\) be as defined in Eq. (5) and let 
\[ c = (\sqrt{\lambda} + R)^2. \]
Then, for \(T \geq 3\),
\[
\frac{1}{T} \sum_{t=1}^{T} f(w_t; A_t) \leq \frac{1}{T} \sum_{t=1}^{T} f(w^*; A_t) + \frac{c \ln(T)}{\lambda T}.
\]

Convergence in obj. function value

Theorem 2. Assume that the conditions stated in Thm. 1 hold and for all \(t\), \(A_t\) is chosen i.i.d. from \(S\). Let \(r\) be an integer picked uniformly at random from \([T]\). Then,
\[
\mathbb{E}_{A_1, \ldots, A_T} \mathbb{E}_r [f(w_r)] \leq f(w^*) + \frac{c \ln(T)}{\lambda T}.
\]
Pegasos Performance

Comparisons to:
- Cutting-plane algorithm (SVM-Perf)
- Decomposition algorithm (SVM-Light)

Table 1. Training time in CPU-seconds

<table>
<thead>
<tr>
<th></th>
<th>Pegasos</th>
<th>SVM-Perf</th>
<th>SVM-Light</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAT</td>
<td>2</td>
<td>77</td>
<td>20,075</td>
</tr>
<tr>
<td>Covertype</td>
<td>6</td>
<td>85</td>
<td>25,514</td>
</tr>
<tr>
<td>astro-ph</td>
<td>2</td>
<td>5</td>
<td>80</td>
</tr>
</tbody>
</table>
Comparisons to SGD

Figure 2. Comparisons of Pegasos to Norma (left) and Pegasos to stochastic gradient descent with a fixed learning rate (right) on the Astro-Physics dataset. In the left plot, the solid lines designate the objective value and the dashed lines depict the loss on the test set.

Norma

- SGD, no projection, \( \eta_t = \frac{p}{\lambda \sqrt{t}}, \ p \in (0, 1) \)

Zhang

- SGD, no projection, \( \eta_t = \eta, \) e.g. \( \eta = 10^{-5} \)
SGD from Last Lecture?

Classical (strongly convex) SGD

**Stepsize:** \( \gamma_t = \frac{\theta}{t}, \quad \theta > \frac{1}{2\sigma} \)

**Convergence in iterates:**
\[
\mathbb{E}[\|x_t - x^*\|_2^2] \leq Q(\theta) \frac{1}{t}
\]
where
\[
Q(\theta) := \max\{\theta^2 M^2 (2\sigma\theta - 1)^{-1}, \|x_1 - x^*\|_2^2\}
\]

**Convergence in obj. function values:**

- When \( f \) is differentiable and has Lipschitz continuous gradients
\[
\mathbb{E}[f(x_t) - f(x^*)] \leq Q(\theta) \frac{L}{2t}
\]
Comparison to SGD

In our next experiment, we compared Pegasos to Norma (Kivinen et al., 2002) and to a variant of stochastic gradient descent described in (Zhang, 2004). Both methods are similar to Pegasos when setting $k = 1$ with two differences. First, there is no projection step. Second, the scheduling of the learning rate, $\eta_t$, is different. In Norma (Thm. 4), it is suggested to set $\eta_t = p/(\lambda \sqrt{t})$, where $p \in (0, 1)$. Based on the bound given in Thm. 4 of (Kivinen et al., 2002), the optimal choice of $p$ is $0.5(2 + 0.5T^{-1/2})^{1/2}$, which for $t \geq 100$ is in the range $[0.7, 0.716]$. Plugging the optimal value of $p$ into Thm. 4 in (Kivinen et al., 2002) yields the bound $O(1/(\lambda \sqrt{T}))$. We therefore hypothesized that Pegasos would converge much faster than the other methods, but due to the lack of space, we omit the graphs.

We now turn to comparing Pegasos to the algorithm from (Zhang, 2004) which simply sets $\eta_t = \eta$, where $\eta$ is a (fixed) small number. A major disadvantage of this approach is that finding an adequate value for $\eta$ is a difficult task on its own. Based on the analysis given in (Zhang, 2004) we started by setting $\eta$ to be $10^{-5}$. Surprisingly, this turned out to be a poor choice and the optimal choice of $\eta$ was substantially larger. In Fig. 5 (right) we illustrate the convergence of stochastic gradient descent with $\eta$ set to be $10^{-5}$.
The Effect of $k$

Figure 3. The effect of $k$ on the objective value of Pegasos on the Astro-Physics dataset. Left: $T$ is fixed. Right: $kT$ is fixed.
Summary

In short, Pegasos algorithm is

- Stochastic subgradient descent
- For strongly convex objective function
- With projection, as in Nemirovski et al., 2009
- Each subgradient is estimated with $k$ examples
- $O(1/t)$ stepsize, which is consistent to Nemirovski
- $O((\ln t)/t)$ convergence in function value
  - Without requiring Lipschitz continuity of gradients
  - Does not depend on $k$
Classical vs. Robust SGD?

Lee & Wright, ASSET: Approximate Stochastic Subgradient Estimation Training for SVMs, 2012

- Also with approximations to nonlinear kernels

Figure 2: Progress of ASSET\(_F\) and ASSET\(_F^*\) to their completion (MNIST-E), in terms of test error rate.
Classical vs. Robust SGD

Table 3: Training CPU time (in seconds, h:hours) and test error rate (%) in parentheses. Kernel approximation dimension is varied by setting $s = 512$ and $s = 1024$ for ASSET$_M$, ASSET$^*_M$, CPSP and CPNY. Decomposition methods do not depend on $s$, so their results are the same in both tables.

<table>
<thead>
<tr>
<th></th>
<th>Subgradient Methods</th>
<th>Cutting-plane</th>
<th>Decomposition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASSET$_M$</td>
<td>ASSET$^*_M$</td>
<td>CPSP</td>
</tr>
<tr>
<td><strong>s = 512</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADULT</td>
<td>23(15.1±0.06)</td>
<td>24(15.1±0.06)</td>
<td>3020(15.2)</td>
</tr>
<tr>
<td>MNIST</td>
<td>97 (4.0±0.05)</td>
<td>101 (4.0±0.04)</td>
<td>550 (2.7)</td>
</tr>
<tr>
<td>CCAT</td>
<td>95 (8.2±0.08)</td>
<td>99 (8.3±0.06)</td>
<td>800 (5.2)</td>
</tr>
<tr>
<td>IJCNN</td>
<td>87 (1.1±0.02)</td>
<td>89 (1.1±0.02)</td>
<td>727 (0.8)</td>
</tr>
<tr>
<td>COVTYPE</td>
<td>697(18.2±0.06)</td>
<td>586(18.2±0.07)</td>
<td>1.8h(17.7)</td>
</tr>
<tr>
<td><strong>s = 1024</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADULT</td>
<td>78(15.1±0.05)</td>
<td>83(15.1±0.04)</td>
<td>3399(15.2)</td>
</tr>
<tr>
<td>MNIST</td>
<td>275 (2.7±0.03)</td>
<td>275 (2.7±0.02)</td>
<td>1273 (2.0)</td>
</tr>
<tr>
<td>CCAT</td>
<td>265 (7.1±0.05)</td>
<td>278 (7.1±0.04)</td>
<td>2950 (5.2)</td>
</tr>
<tr>
<td>IJCNN</td>
<td>307 (0.8±0.02)</td>
<td>297 (0.8±0.01)</td>
<td>1649 (0.8)</td>
</tr>
<tr>
<td>COVTYPE</td>
<td>2259(16.5±0.04)</td>
<td>2064(16.5±0.06)</td>
<td>4.1h(16.6)</td>
</tr>
</tbody>
</table>

Robust   Classical