Section 06: Solutions

1. Stochastic Gradient Descent

Consider minimizing an average of functions:

$$\min_w \frac{1}{n}\sum_{i=1}^{n} \ell_i(w),$$

where w is a $d-$dimensional vector (or the feature dimension is d). The minimization of the negative of a log-likelihood function can serve as an example. Recall that the (full) gradient descent step is given by

$$w^{(t+1)} = w^{(t)} - \eta \cdot \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i\big(w^{(t)}\big).$$

The computational cost of a single step here is $\mathcal{O}(dn)$. To reduce cost, one idea is to just use a subset of all samples to approximate the full gradient. Specifically, consider revising the gradient descent step as follows:

$$w^{(t+1)} = w^{(t)} - \eta \cdot \nabla \ell_{I_t}(w^{(t)}),$$

where I_t is chosen randomly within $\{1, 2, ..., n\}$ with equal probabilities. This is called **stochastic gradient descent (SGD)**, and the computational cost of a single step now reduces to $\mathcal{O}(d)$.

(a) The following two results provide intuitions or foundations for why SGD works.

- $\mathbb{E}_{I_t}(\nabla \ell_{I_t}(w^{(t)})) = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i\big(w^{(t)}\big)$, which is the full gradient. Hence the estimate of gradient is unbiased.
- Let $\ell(w) = \frac{1}{n}\sum_i \ell_i(w)$ and $w^* = \arg\min_w \ell(w)$. Assume $||w^{(1)} - w^*||_2^2 \le R$ and $\sup_w \max_i ||\nabla \ell_i(w)||_2^2 \le G$. Then

  $$\mathbb{E}[\ell(\bar{w}) - \ell(w^*)] \le \sqrt{\frac{RG}{T}},$$

  where $\bar{w} := \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$. Therefore, the expected error over T iterations is $\mathcal{O}(\frac{1}{\sqrt{T}})$. (The proof of this result is provided in the solution part for reference.)

Solution:

We first consider deriving an upper bound for $\mathbb{E}(\ell(w^{(t)}) - \ell(w^*))$:

$$\begin{aligned}
\mathbb{E}||w^{(t+1)} - w^*||_2^2 &= \mathbb{E}||w^{(t)} - \eta \nabla \ell_{I_t}(w^{(t)}) - w^*||_2^2 \\
&= \mathbb{E}||w^{(t)} - w^*||_2^2 - 2\eta \mathbb{E}\Big(\nabla \ell_{I_t}(w^{(t)})^T (w^{(t)} - w^*)\Big) + \eta^2 \mathbb{E}||\nabla \ell_{I_t}(w^{(t)})||_2^2 \\
&\le \mathbb{E}||w^{(t)} - w^*||_2^2 - 2\eta \mathbb{E}\Big(\nabla \ell_{I_t}(w^{(t)})^T (w^{(t)} - w^*)\Big) + \eta^2 G \\
&\le \mathbb{E}||w^{(t)} - w^*||_2^2 - 2\eta \mathbb{E}(\ell(w^{(t)}) - \ell(w^*)) + \eta^2 G,
\end{aligned}$$

where the last inequality holds because of the following:

$$\begin{aligned}
\mathbb{E}\Big(\nabla \ell_{I_t}(w^{(t)})^T (w^{(t)} - w^*)\Big) &= \mathbb{E}\mathbb{E}\Big[\nabla \ell_{I_t}(w^{(t)})^T (w^{(t)} - w^*) | I_1, w^{(1)}, ..., I_{t-1}, w^{(t-1)}\Big] \\
&= \mathbb{E}\frac{1}{n}\sum_i \nabla \ell_i(w^{(t)})^T (w^{(t)} - w^*) \\
&= \mathbb{E}\nabla \ell(w^{(t)})^T (w^{(t)} - w^*) \\
&\ge \mathbb{E}\Big(\ell(w^{(t)}) - \ell(w^*)\Big),
\end{aligned}$$

where the last inequality holds from the convexity of $\ell(\cdot)$. Therefore, we've proved that $\mathbb{E}||w^{(t+1)} - w^*||_2^2 \le$

$\mathbb{E}||w^{(t)} - w^*||_2^2 - 2\eta\mathbb{E}(\ell(w^{(t)}) - \ell(w^*)) + \eta^2 G$, which implies (from rearrangement) that

$$\mathbb{E}(\ell(w^{(t)}) - \ell(w^*)) \leq \frac{1}{2\eta}\Big(\mathbb{E}||w^{(t)} - w^*||_2^2 - \mathbb{E}||w^{(t+1)} - w^*||_2^2 + \eta^2 G\Big). \tag{1}$$

Now note that the convexity of ℓ and Jensen's inequality ensure that $\ell(\bar{w}) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(w^{(t)})$, which implies

$$\mathbb{E}\Big(\ell(\bar{w}) - \ell(w^*)\Big) \leq \frac{1}{T}\sum_t \mathbb{E}\Big(\ell(w^{(t)}) - \ell(w^*)\Big). \tag{2}$$

From (**??**) and (**??**), we have

$$\begin{aligned}
\mathbb{E}\Big(\ell(\bar{w}) - \ell(w^*)\Big) &\leq \frac{1}{T}\sum_t \mathbb{E}\Big(\ell(w^{(t)}) - \ell(w^*)\Big) \\
&\leq \frac{1}{T}\sum_t \frac{1}{2\eta}\Big(\mathbb{E}||w^{(t)} - w^*||_2^2 - \mathbb{E}||w^{(t+1)} - w^*||_2^2 + \eta^2 G\Big) \\
&= \frac{1}{2\eta T}\Big(\mathbb{E}||w^{(1)} - w^*||_2^2 - \mathbb{E}||w^{(T+1)} - w^*||_2^2\Big) + \frac{\eta G}{2} \\
&\leq \frac{1}{2\eta T}\mathbb{E}||w^{(1)} - w^*||_2^2 + \frac{\eta G}{2} \\
&\leq \frac{R}{2\eta T} + \frac{\eta G}{2} \\
&= \sqrt{\frac{RG}{T}},
\end{aligned}$$

where the last equality holds by choosing $\eta = \sqrt{\frac{R}{GT}}$.

(b) What disadvantages can SGD have? How can we balance between the noise in updates and computational cost?

Solution:

By treating SGD as noise-injected gradient descent:

$$\nabla\ell_{I_t}(w^{(t)}) = \mathbb{E}_{I_t}[\nabla\ell_{I_t}(w^{(t)})] + e_t = \frac{1}{n}\sum_{i=1}^{n}\ell_i(w^{(t)}) + e_t,$$

where e_t represents the noise term and is random, we know that the steps taken towards a minimum can be very noisy because the gradient used in updating involves noise. One way to balance the noise in updates and computational cost is to consider a technique called **mini-batching**, which is employed with SGD.

(c) Gradient descent requires the full gradient when updating while (standard) SGD utilizes the gradient of one sample when updating. **Mini-batching** is somewhere between the two extremes. That is, we choose a random subset $I_t \subseteq \{1, ..., n\}$ with size $|I_t| = b \ll n$ in the stochastic gradient descent step:

$$w^{(t+1)} = w^{(t)} - \eta \cdot \frac{1}{b}\sum_{i_t \in I_t}\nabla\ell_{i_t}(w^{(t)}).$$

With mini-batching, we have the following results:

- $\mathbb{E}_{I_t}\Big(\frac{1}{b}\sum_{i_t\in I_t}\nabla\ell_{i_t}(w^{(t)})\Big) = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell_i(w^{(t)})$: we still have an unbiased estimate of the full gradient.
- Compared to standard SGD, variance of the gradient estimate is reduced approximately by $\frac{1}{b}$.

- Computational cost for each step now becomes $\mathcal{O}(db)$.

Remark: By matrix computations (computing b gradients at a time) and parallelization, we can denoise the estimated gradients without increasing much computational cost (for batch size b that is not large).

(d) How should we choose the batch size? Are there other extensions or variants of the basic stochastic gradient descent algorithm?

Solution:

The choice of the optimal batch size is not an easy question, and there is no standard answer to it. However, we still try to provide some important intuitions regarding the choice of batch size. Firstly, when the objective function (to be minimized) behaves "better" (e.g., Lipschitz continuous, strong convex) than convex functions, the difference in the convergence rates between GD and SGD becomes significant, suggesting a nontrivial gain of having a faster convergence rate and hence we should consider relatively larger batch size. Secondly, a smaller batch size yields less stable gradient estimates, suggesting that we shall employ a fairly small step size/learning rate. An increase in the batch size can be paired with an increase in the step size/learning rate.

Many improvements, which are listed below, on the basic SGD algorithm have been developed and used.

- Implicit updates (ISGD)
- Momentum
- Averaged stochastic gradient descent
- Adaptive gradient algorithm (AdaGrad)
- Root Mean Square Propagation (RMSProp)
- Adaptive Moment Estimation (Adam)

Basically, these methods consider to fine-tune the step size parameter, take previous update magnitude into account, or introduce the second moments of the gradients when updating. For example, Momentum remembers the previous update magnitude so that $w^{(t)}$ tends to keep traveling in the same direction, preventing oscillations:

$$w^{(t+1)} = w^{(t)} - \eta \nabla \ell_{I_t}(w^{(t)}) + \alpha(w^{(t)} - w^{(t-1)}).$$

Adam, as another example, considers to tune to step size with the second moments of the gradients:

$$w^{(t+1)} = w^{(t)} - \eta G\Big(\nabla \ell^{(t)}, \nabla \ell^{(t-1)}, ..., (\nabla \ell^{(t)})^2, (\nabla \ell^{(t-1)})^2, ...\Big),$$

where $\nabla \ell^{(t)} = \nabla \ell_{I_t}(w^{(t)})$ and G is a function that involves element-wise square of all previous gradients. The paper below provides more details on Adam:

https://arxiv.org/pdf/1412.6980.pdf

2. SVMs

Consider the dataset consisting of 7 data points, 4 with positive labels $\{0, 1, 2, 3\}$, and 3 with negative labels $\{-3, -2, -1\}$. Suppose we want to learn a linear SVM with slack variables for this dataset. Recall we can for-

malize this as a constrained optimization problem:

$$\min_{w,b,\xi} ||w||^2 + C\sum_i \xi_i$$

$$\text{subject to}$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \; \forall i$$

$$\xi_i \geq 0 \; \forall i$$

where C is a regularization parameter that balances the size of the margin (smaller $||w||^2$) vs. the violation of the margin (smaller $\sum_i \xi_i$).

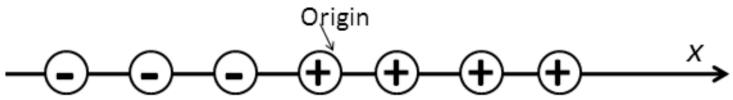

Figure 1: Dataset

Also recall that a support vector is a datapoint that lies on a margin.

(a) If $C = 0$, which means we only care about the size of the margin, how many support vectors do we have? **Solution:**

7, if we just want to minimize $||w||^2$, then we will vary b and ξ_i so all of the constraints are satisfied. As $||w|| \to 0$, we only have to satisfy $y_i b \geq 1 - \xi_i$ and $\xi_i \geq 0 \; \forall i$. These constraints can all be satisfied with $\xi_i \in \{0, 2\}$ and $b = 1$, letting all of the vectors be support vectors.

(b) If $C \to \infty$, which means we only care about the violation of the margin, how many support vectors do we have? **Solution:**

2, if $C \to \infty$, then we just care about minimizing $\sum_i \xi_i$. That is done when we have only one margin between the two classes at -0.5 with two defining support vectors -1 and 0.