Contents

1 Stability in stochastic optimization

Here we summarize some of the key ideas and techniques from [1, 2]. The motivation for seeking "better models" in stochastic optimization is to reduce the impact and burden of hyperparameter selection, in particular, selection of the initial learning rate and learning rate schedule. Without proper selection of the learning rate, stochastic gradient descent can diverge or converge arbitrarily slowly[1], even on convex problems.

Somewhat more surprisingly, [1] shows fast convergence on a class of so-called "easy to optimize" problems (including convergence to stationary points for weakly-convex problems). As an application, they show that their methods give the fastest rates known for phase retrieval.

We also relate the work on stability and convergence guarantees to those studied in the context of gradient clipping [3, 4] and empirical risk minimization for non-convex objectives [5]. The end of the notes also includes experimental results demonstrating the improved stability and convergence rates on synthetic and real datasets[2].

1.1 Instability of gradient descent

Gradient descent can behave badly even for the quadratic objective $F(x) = 1/2x^2$. For a learning rate schedule: $\alpha_k = \alpha_0 k^{-\beta}$ gradient descent proceeds as: $x_{k+1} = (1 - \alpha_0 k^{-\beta})x_k$. We have $|x_{k+1}| = |(1 - \alpha_0 k^{-\beta})||x_k|$ which for poorly specified α_0 can be lower bounded

[1] I haven't gone through this example, but Asi and Duchi [2] cite Nemirovski for this [6]

[2] Code available on github.com/dmh43/research

by $2|x_k| = O(2^{k-1})$ for all $k \leq k_0$ where k_0 can potentially be large. So even in the case of a quadratic objective, the learning rate selection can lead to poor convergence[3].

The story is worse for more "difficult" objectives such as $F(x) = (e^x + e^{-x})$ which diverges for all polynomial learning rate schedules if the initialization is large enough[4].

1.2 Problem setup and methods

Consider the optimization problem:

$$\min_{x \in \mathcal{X}} F(x)$$

over a convex set $\mathcal{X}$ where $F(x) = \mathbb{E}_P f(x; S)$ for some loss function f parametrized by x and S is a sample distributed as P. Given a sequence of samples $S_k \sim P$, The stochastic gradient method (SGM) proceeds by iterating:

$$x_{k+1} = x_k - \alpha_k g_k$$

where g_k is an element of the subgradient of $f(x_k; \cdot)$ at S_k. Presenting SGM in this way frames it as a noisy approximation of gradient descent. Alternatively, we can view SGM as a minimization of a sequence of (random) linear approximations to the objective function f. To that end, consider the linear approximation to f at x_k:

$$f_{x_k}^{\text{SGM}}(y; s) = f(x_k; s) + \langle g_k, y - x_k \rangle$$

Now we can frame SGM as minimizing a regularized version of this objective, iterating:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} f_{x_k}^{\text{SGM}}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2$$

Proximal point and prox-linear methods can also be viewed in this model-based framework. For the duration of these notes, we focus on the so-called *truncated model* which applies when f has a known lower bound. Without loss of generality, we assume this lower bound is 0: $f(\cdot; s) > 0$ for all s in the sample space $\mathcal{S}$:

$$f_x^{\text{trunc}}(y; s) = \left[f_x^{\text{SGM}}(y; s) \right]_+ = \max\{0, f(x; s) + \langle g_k, y - x \rangle\}$$

This gives us the truncated model method:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} f_{x_k}^{\text{trunc}}(x_k; s) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \tag{1}$$

$$= \operatorname*{argmin}_{x \in \mathcal{X}} \max\{0, f(x_k; s) + \langle g_k, x - x_k \rangle\} + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \tag{2}$$

[3] See example 2 in [2]

[4] See this math stack exchange question

1.3 From better models to stability to (asymptotically) optimal convergence

Arguably, the truncated model of the previous section is "better" (or at least more faithful) than the naive linear model of SGM. Asi and Duchi show [1] that for convex functions whose gradients grow at most polynomially, the iterates of the truncated method are bounded with probability 1. This is precisely the additional condition needed for proposition 1 and theorem 2 which together guarantee almost sure convergence of the iterates to the optimum as well as asymptotically optimal convergence.

In particular, the averaged iterates $\bar{x}_k = 1/k \sum_i^k x_i$ are asymptotically normal:

$$\sqrt{k}(\bar{x}_k - x^*) \xrightarrow{d} N(0, \Sigma(x^*))$$

where $\Sigma(x^*) = \nabla^2 F(x^*)^{-1} Cov(\nabla f(x^*; S)) \nabla^2 F(x^*)^{-1}$.

1.3.1 Local Asymptotic Minimax Optimality

As in the context of estimation [9 section 8.3], minimax lower bounds are often too coarse (see the discussion in the introduction of [10]), so we instead consider minimax optimallity within a neighborhood that shrinks as the iteration number increases (for M-estimation we consider smaller parameter sets with increasing sample size). Generally, in stochastic optimization, the objective function is known, but the distribution of the samples $S \sim P$ is not known. Analogously to the estimation case, we consider a shrinking neighborhood of distributions (with KL divergence within some range).

Corollary 2 of [10] gives us the following result (recalled informally): the limit behavior of the worst case $\|\sqrt{k}(\bar{x}_k - x^*)\|_2$ is lower bounded by $\|Z\|_2$ where $Z \sim N(0, \Sigma(x^*))$. This shows that the convergence of the truncated method is locally asymptotically optimal in the minimax sense. As in the estimation case (recall the behavior of the Hodges estimator at 0), it can be shown that the set of samples on which $\bar{x}_k$ is strictly better than Z in expectation has measure 0.

1.3.2 Results for weakly-convex functions

Going beyond the results for convex functions developed in [2], Asi and Duchi also show that for coercive[5], weakly convex functions that satisfy the following "relative noise" condition on f', the proximal model method has bounded iterates [1]:

$$Var(f'(x; S)) \leq C_1 \|F'(x)\|_2^2 + C_2$$

Although there is no discussion of this condition for the truncated model method, the condition looks very similar to the relaxation of the Lipschitz smoothness condition explored in [3] as part of their analysis of gradient clipping and is also a relaxation of the condition of globally Lipschitz gradients. Accordingly, we include a comparison to SGM with gradient clipping in our experimental results section. Recall that for a function to have L-Lipschitz gradients it must satisfy:

[5] Go to infinity as the norm of the argument goes to infinity

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \text{ for all } x, y$$

Even a simple third order polynomial does not satisfy this condition over all $\mathbb{R}^d$. To relax this condition, [3] introduces the notion of (L_0, L_1) smoothness for twice differentiable functions as:

$$\|\nabla^2 f(x)\|_2 \leq L_0 + L_1\|\nabla f(x)\|_2$$

which has some similarity to the "relative noise" condition above.

In addition, for weakly-convex functions whose set of stationary points have an image of Lebesgue measure 0, [1] shows that methods with bounded iterates converge to stationary points. This is reminiscent of the conditions and results in [5][6]. In an attempt to connect the two, we check the required conditions for an application addressed in [5] (non-convex binary classification) as well as explore some experimental results for this problem.

1.4 Fast rates for easy problems

Asi and Duchi [1, 2] call a problem *easy to optimize* if the loss of a optimum, x^*, at a specific sample, s, is the best possible loss achievable over all $x \in \mathcal{X}$ for every s in the sample space[7]:

$$\inf_{x \in \mathcal{X}} f(x; s) \;\; f(x^*; s) \, \forall s \in \mathcal{S}$$

While certainly a strong assumption, problems that are easy to optimize include: phase retrieval, classification problems with linearly separable classes, and machine learning problems where the training loss goes to 0 (or it's absolute minimum). In the experimental section we also consider the problem of solving an overdetermined system (specifically, noiseless ℓ_1 regression).

For easy to optimize problems, Lemma 4.1 of [1] shows that the sequence $\|x_k - x^*\|_2$ is non-increasing for the truncated model method (as well as the other methods considered in [1]). In order to derive rates of convergence, we also require an additional "sharp-growth" assumption near the optimum set $\mathcal{X}^*$ (A6 from [1]) that is awkward to work with. Instead we go after a pair of simpler conditions that imply the sharp growth condition:

The first is a small-ball type condition [7] where we require the following lower bound for some constants $\lambda, p > 0$[8, 9]:

$$\mathbb{P}\left(f(x; S) \geq \lambda \text{dist}(x, \mathcal{X}^*)\right) \geq p$$

The second is that the gradients grow at most quadratically:

[6] We went through this paper a few weeks ago.

[7] Note that this condition implies that $\inf_{x \in \mathcal{X}} f(x; s)$ exists.

[8] [2] mentions that an estimate of this type can be obtained from an application of the Paley-Zygmund inequality, but it's not clear to me that it has been demonstrated in this work or in [1].

[9] Recall that we assume wlog that the infimum of f is 0.

$$\mathbb{E}\|f'(x;S)\|_2^2 \leq C(1+\text{dist}(x,\mathcal{X}^*))^2$$

With these conditions, (a more specific version of) proposition 2 of [1] gives us:

For stepsizes of the form $\alpha_k = \alpha_0 k^{-\beta}$ where $\beta < 1$, the iterates of the truncated model method for a convex objective converge linearly to the optimum with probability 1. More specifically:

$$\frac{\text{dist}(x_k,\mathcal{X}^*)}{(1-\lambda_1)^k} \xrightarrow{a.s.} V$$

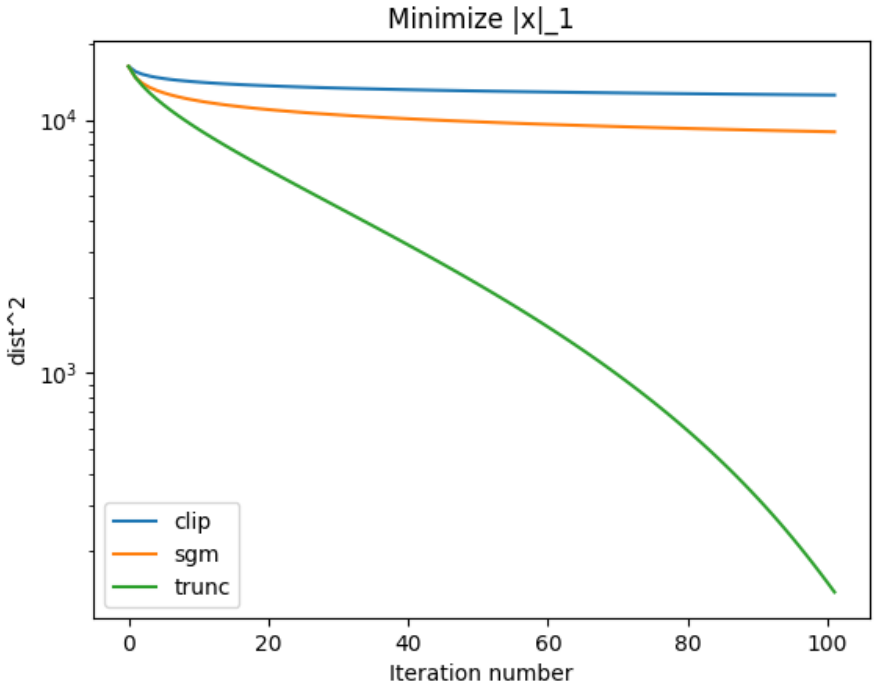
For some finite limit V where λ_1 is a constant related to the sharp growth condition mentioned previously. In contrast, consider the objective $f(x) = \|x\|_1$ (which satisfies the conditions above for all distributions[10]) where the convergence rate of subgradient descent methods is bounded by that of α_k. Note that in this case SGM iterates:

$$x_{k+1} = x_k - \alpha_k \text{sign}(x_k)$$

So the the learning rate directly determines the convergence rate in the SGM case. Recall that a typical learning rate schedule is $\alpha_k = O(1/\sqrt{k})$ [12] which is certainly slower than linear convergence. This is the disappointing behavior of the generic subgradient method which does not take advantage of additional information about the objective function[11]:

[10] Note that the gradients are uniformly bounded from above and that $\mathbb{P}\left(\|x\|_1 \geq \epsilon\|x\|_2\right) = 1$ for $\epsilon = \sqrt{n}$.

[11] The *clip* method will always do worse than SGM in this case since the gradient is constant in magnitude throughout (except at the optimum).



For comparison, FISTA (which does not take advantage of the boundedness and *easy to optimize* behavior) achieves convergence of $O(1/k^2)$ [12].

The proof of proposition 2 and it's lemmas involve multiple applications of the Robbins-Siegmund almost supermartingale convergence theorem (Lemma A.4 in [1])

With some additional assumptions on the distribution of the design matrix, example 4 of [1] shows that these conditions hold for phase retrieval. Their derivation involves a typical VC bound on the deviation of the empirical counts from it's expectation [11 Theorem 12.5].

1.5 Experimental Results

We now investigate the performance of the truncated model method in comparison to SGM and SGM with gradient clipping. Although there are more sophisticated methods for implementing gradient clipping such as coordinate-wise adaptive clipping [4] we simply implement clipping to limit the norm of the gradient step. Since clipping is not our focus, we choose a fixed clipping threshold ($\gamma = 0.25$) that was chosen from a range and seemed to perform well in practice (typically better than standard SGM). The clipped SGM update is computed as:

$$h_k = \alpha_k \min\{1, \frac{\gamma}{\|g_k\|_2}\} \tag{3}$$

$$x_{k+1} = x_k - h_k g_k \tag{4}$$

We explore the easy to optimize problem of noiseless ℓ_1 regression (essentially solving a system of linear equations) in addition to the ℓ_2 regression for linear binary classification with sigmoid activation. As we will see, the latter problem is not easy to optimize but we still get some benefit over SGM and clipped SGM.

1.5.1 Truncated model method step

Recall the iteration of the truncated model method:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \max\{0, f(x_k; s) + \langle g_k, x - x_k \rangle\} + \frac{1}{2\alpha_k}\|x - x_k\|_2^2$$

In order to determine a closed form iteration, we'll consider this minimization under $\mathcal{X} = \mathbb{R}^d$ (unconstrained). First reparametrize the above problem in terms of $u = x - x_k$:

$$\min_u \max\{0, f_k + \langle g_k, u \rangle\} + \frac{1}{2\alpha_k}\|u\|_2^2$$

where we've written f_k in place of $f(x_k; s)$. Next, realize that to optimize the above expression u must be in the direction of $-g_k$ so we can further simplify the problem to a 1 dimensional optimization:

$$\min_\lambda \max\{0, f_k - \lambda\|g_k\|_2^2\} + \frac{\lambda^2}{2\alpha_k}\|g_k\|_2^2$$

For $f_k - \lambda\|g_k\|_2^2 \geq 0$ we have that $\lambda = \alpha_k$. Otherwise, λ is determined by solvingd the constrained optmization problem:

$$\begin{aligned} \min_\lambda \frac{\lambda^2}{2\alpha_k}\|g_k\|_2^2 \\ \text{subject to } f_k - \lambda\|g_k\|_2^2 < 0 \end{aligned}$$

After dualizing the constraint and applying the KKT conditions, we get that $\lambda = \frac{f_k}{\|g_k\|_2^2}$. Summarizing, the truncated model method iterates as:

$$\begin{aligned} \lambda_k &= \min\{\alpha, \frac{f_k}{\|g_k\|_2^2}\} \\ x_{k+1} &= x_k - \lambda_k g_k \end{aligned}$$

1.5.2 Step sizes

We choose step sizes of the form $\alpha_0 k^{-\beta}$ for all methods considered. For the truncated method we choose $\beta \approx 1$, and for SGM and the clipped SGM method $\beta = 1$.

1.5.3 Noiseless ℓ_1 Regression

Consider the overspecified system of equations:

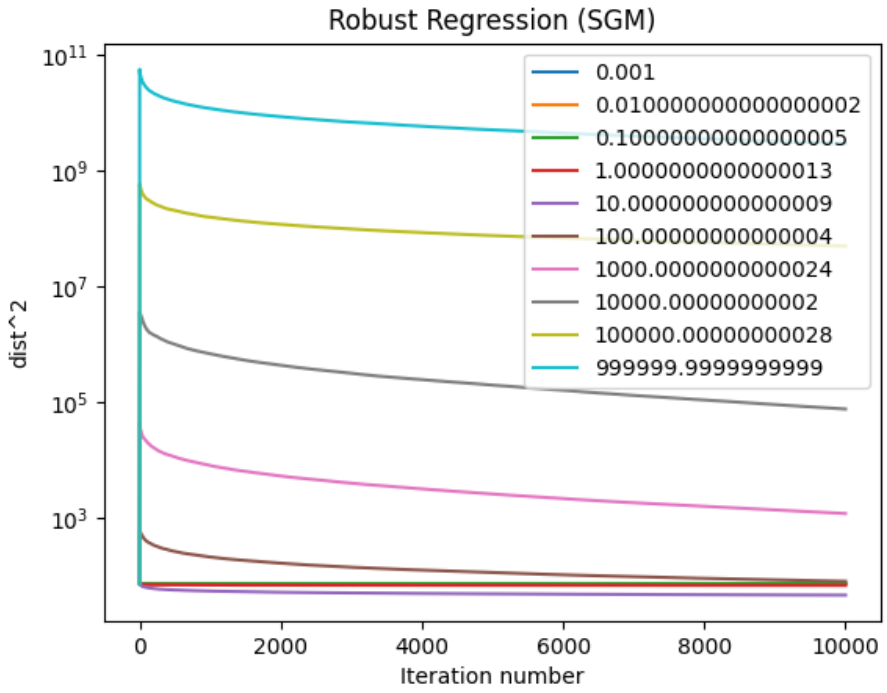
$$y = Ax_0$$

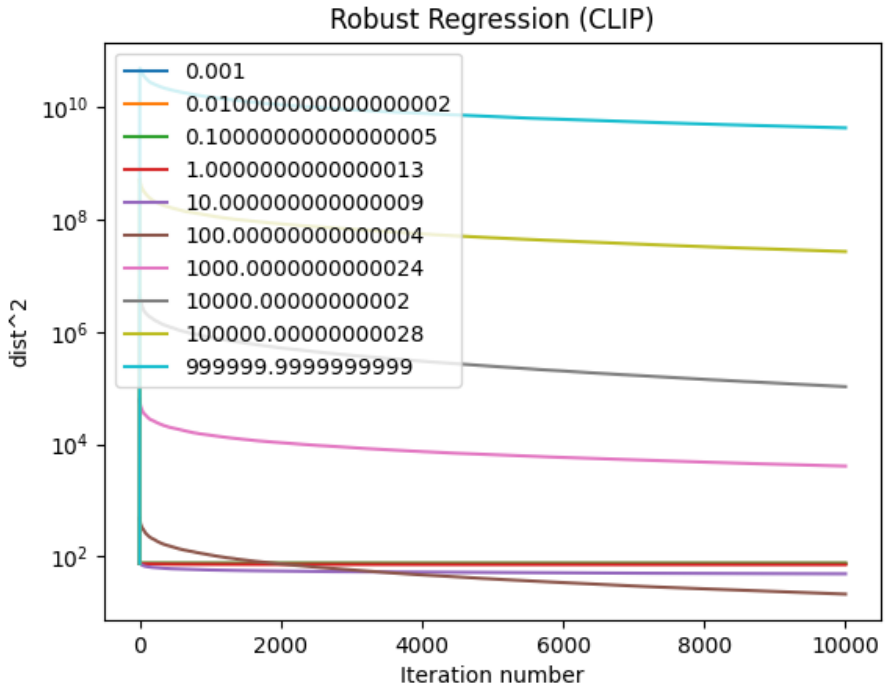
where $y \in \mathbb{R}^n, A \in \mathbb{R}^{n \times d}, x_0 \in \mathbb{R}^d$ where $n > d$. To determine x_0, we'll solve the following unconstrained optimization problem where the expectation is taken over the indexes \$i $\sim$ Uni\1...n\:

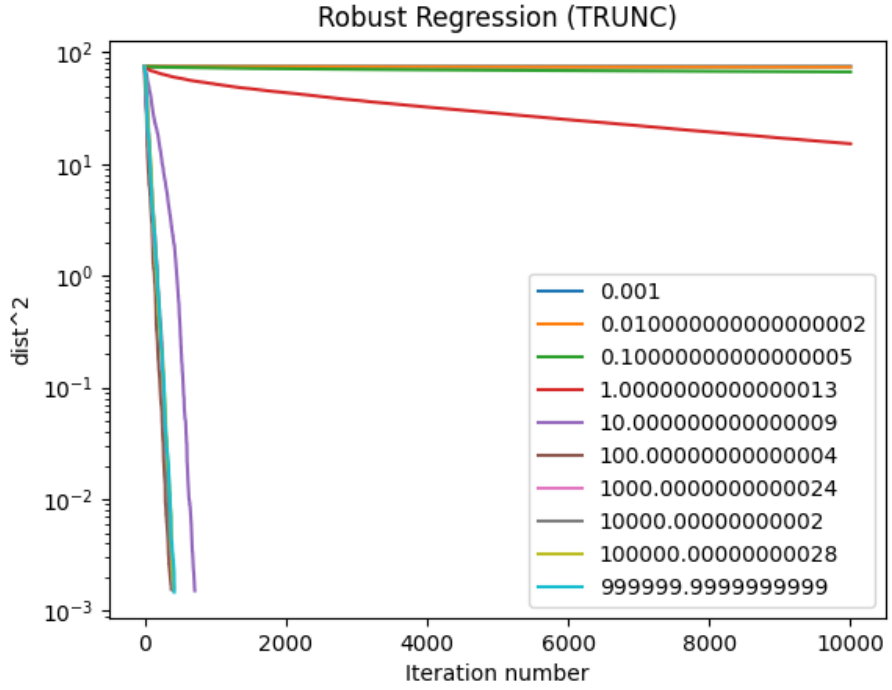
$$\min_x \mathbb{E} \left| y_i - a_i^T x \right|$$

This problem is clearly easy to optimize since the global optimum is 0 and a single sample can achieve 0 error. To generate a problem that satisfies the required sharp growth assumption, we'll sample A from the Haar distribution over orthonormal matrices using `scipy.stats.ortho_group.rvs`. We compare the three methods by plotting the distance from the optimum over a wide range of logarithmically spaced learning rates. Notice that the SGM and clipped SGM methods clearly show their slow convergece of $1/\sqrt{k}$, although the clipped method appears to be slightly less sensitive to stepsize selection. The truncated model method is better behaved over the range of stepsizes and show's clearly the linear convergence.

Although not shown here, experiments using a poorly conditioned A matrix make the difference in the methods even more dramatic. In contrast, if we instead considered ℓ_1 regression with noise (which is not easy to optimize), we would see that the methods perform similarly.


Robust Regression (SGM)
dist^2
Iteration number
0.001
0.010000000000000002
0.10000000000000005
1.0000000000000013
10.000000000000009
100.00000000000004
1000.0000000000024
10000.00000000002
100000.00000000028
999999.9999999999


Robust Regression (CLIP)
dist^2
Iteration number
0.001
0.010000000000000002
0.10000000000000005
1.0000000000000013
10.000000000000009
100.00000000000004
1000.0000000000024
10000.00000000002
100000.00000000028
999999.9999999999



1.5.4 Non-convex Linear Binary Classification

The second problem we consider is a non-convex alternative to logistic regression where the prediction is of the form $\sigma(x^T a)$ where $\sigma(\cdot)$ is the logistic sigmoid function. The optimization problem at hand is given by:

$$\min_{x \in \mathcal{X}} \mathbb{E}\left(y_i - \sigma(x^T a_i)\right)^2$$

Which is certainly not easy to optimize in general; the optimum x for a single sample does not exist (x will go off to ∞ or $-\infty$ to minimize the loss). Further, the objective has globally Lipschitz gradients, so we expect the methods to behave similarly with regards to convergence rates[12]. However we might still benefit from increased stability over a wider range of stepsizes.

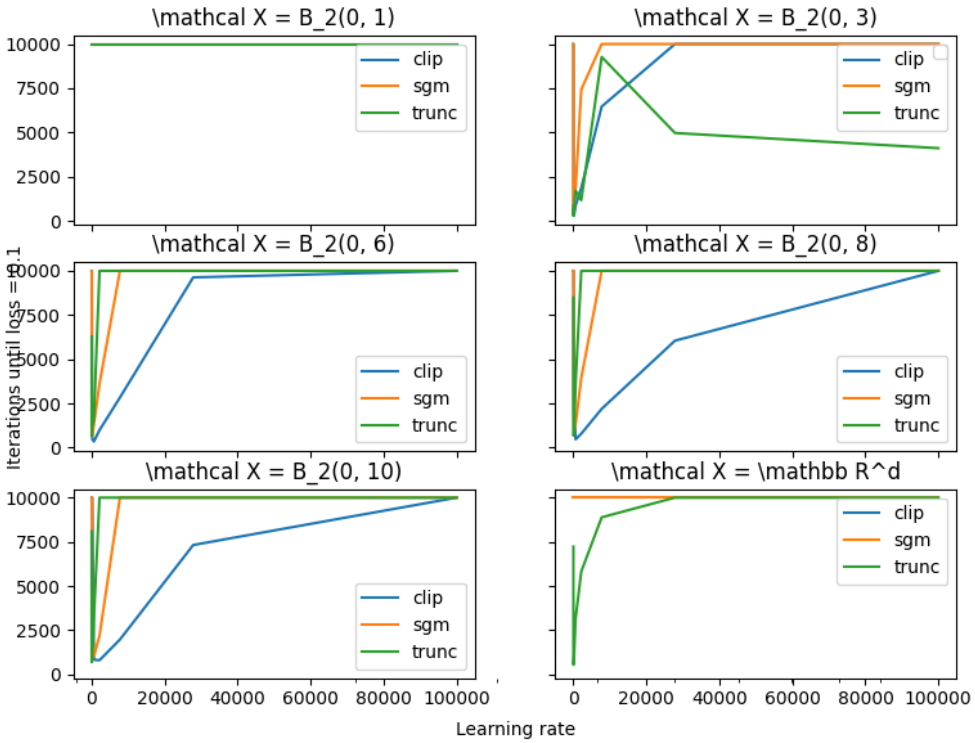
This optimization problem was studied by Mei et al. in [5] as an application of their results on the convergence of critical points of empirical risk for non-convex objectives. In particular, they consider the optimization with constraint set $\mathcal{X} = B(0, r)$ the ℓ_2 ball of radius r centered at the origin. In the case of linearly separable classes, x goes off to infinity, so enforcing some constraint or adding some regularization is typically required.

[12] I dont have a reference for this, but there is a comment at the beginning of the experiments section of [2] that mentions that SGM is asymptotically normal with optimal covariance for objectives with globally Lipschitz gradients, in which case we expect the asymptotic rates of SGM and the truncated model method to be similar. Most likely [8] has some specific references for this (likely some work from Polyak or Shapiro).

Enforcing this hard constraint automatically insures the stability of iterates, so we can directly apply the results of 3.3 from [1] to show that we have convergence to a stationary point. It is easy to show that the objective is weakly convex by computing the hessian and noticing that it is bounded from below. Mei et al. show further that the above problem has only a single stationary point (if r is large enough), and that that point is the global optimum. Proposition 1 of [1] tells us that x_k converges to a stationary point; since there is only a single stationary point, x_k converges to the global optimum.

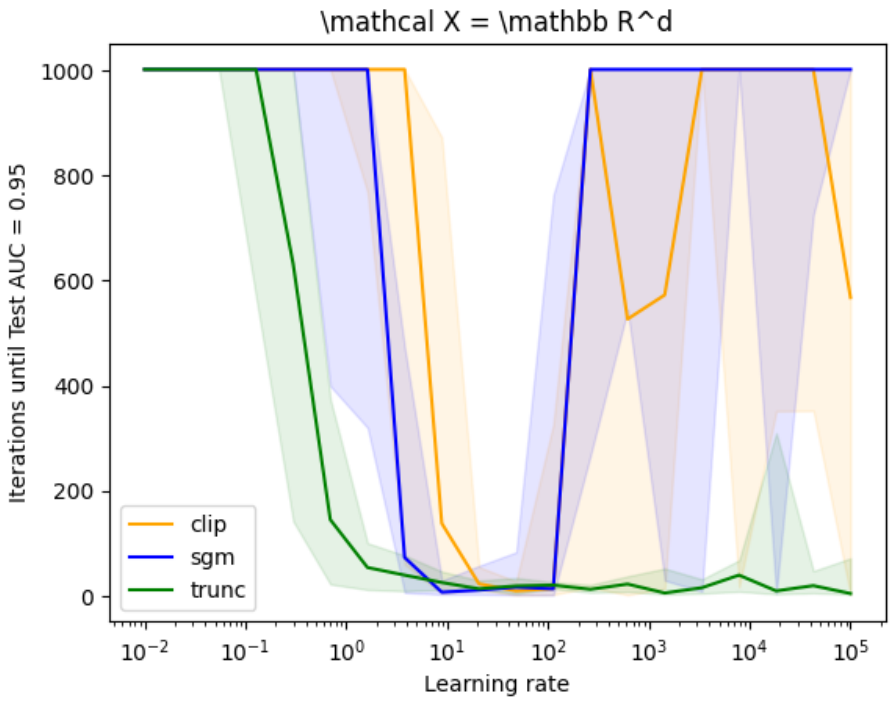
We create a synthetic problem by sampling A from a unit normal and sampling the target, y, as: $y_i = 1\,wp\,\sigma(a_i^T x)$ and 0 otherwise. We sweep over a range of initial step sizes for SGM, clipped SGM and the truncated model method for a range of values for r and plot the number of iterations required to reach a specified error (with a maximum number of iterations of 10^4).

For the finite values of r it is important to exactly solve the iteration specified by the truncated model method, rather than iterating as for the unconstrained problem and then projecting back on to the constraint set. Despite this, we only implemented the latter projected gradient method so our results for finite r might not be representative. Notice that the SGM with clipping performs best for finite r while the truncated model method is the only method to reach the required error within the maximum number of step sizes for the unconstrained problem.



As an example application to real data, we also experimented with the binary classification breast cancer dataset available through sklearn. We only checked the unconstrained

case but ran several bootstrapped samples to obtain 25th and 75th percentiles (plotted in lighter colors). Our stopping criteria in this case was achieving a given validation AUC. Notice that the truncated model method performs well over most of the range of learning rates. All methods are able to achieve a test AUC of 0.99 for some initial step size and value of r. Interestingly, logistic regression could not achieve a test AUC much better than 0.95.



2 Questions

- The paragraph in Example 4.3.1 from [2] following lemma 4.2 includes the following claim for $a/\sqrt{n} \sim \text{Uni}[\mathbb{S}^{n-1}]$ uniformly distributed on the sphere of radius $\sqrt{n}$ in n dimensions and any $v \in \mathbb{R}^n$:

$$\mathbb{P}\left(\langle a, v\rangle \geq \frac{1}{2}\|v\|_2\right) \geq \frac{1}{2}$$

See this stack overflow question: Lower bound on surface area of hyperspherical cap of height $O(1/\sqrt{n})$

- Where is the last inequality from example 4 of [1] from? Are some of the terms being uniformly bounded by a constant?

- What is the connection between the truncated model method and Polyak step sizes? How is this related to gradient clipping (in a deeper way than addressed here)?

3 References

[1] Asi, H. & Duchi, J. C. The importance of better models in stochastic optimization. Arxiv (2019).
[2] Asi, H. & Duchi, J. C. Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity. Arxiv (2018) doi:10.1137/18m1230323.
[3] Zhang, J., He, T., Sra, S. & Jadbabaie, A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. Arxiv (2019).
[4] Sra, S. Why Adam Beats SGD for Attention Models. (n.d.).
[5] Mei, S., Bai, Y. & Montanari, A. The Landscape of Empirical Risk for Non-convex Losses. Arxiv (2016).
[6] Nemirovski, A., Juditsky, A., Lan, G. & Shapiro, A. Robust Stochastic Approximation Approach to Stochastic Programming. Siam J Optimiz 19, 1574–1609 (2009).
[7] Mendelson, S. Learning without Concentration. Arxiv (2014).
[8] Duchi, J. & Ruan, F. Asymptotic Optimality in Stochastic Optimization. Arxiv (2016).
[9] Vaart, van der. Asymptotic Statistics. 1–458 (1998).
[10] Duchi, J. & Ruan, F. Asymptotic Optimality in Stochastic Optimization. Arxiv (2016).
[11] Devroye, L., Gyorfi, L. & Lugosi, G. A Probabilistic Theory of Pattern Recognition. Discrete Appl Math 73, 192–194 (1997).
[12] Bubeck, S. Convex Optimization: Algorithms and Complexity. Found Trends Mach Learn 8, 231–357 (2015).