Project 1: Report

Justin Baker, Eric Brown, Trent DeGiovanni,
Edward Gu, Rebecca Hardenbrook

October 15, 2021

Consider training the following regularized logistic regression model

$$\min_x F(x) := f(x) + \lambda R(x)$$

where

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T x)),$$

with n being the sample size and $a_i \in \mathbb{R}^d$ ($d = 50$) is the training data, $b_i \in \{-1, 1\}$ be the label of a_i. Here, we consider two different regularization functions i.e. ℓ_1-regularization ($R(x) = ||x||_1$) and ℓ_2-regularization ($R(x) = ||x||_2^2$).
Please use the code in the zip file to generate 1000 data-label pairs $\{a_i, b_i\}_{i=1}^{1000}$.

1. Derive $\text{prox}_{\lambda||x||_1}\{f(x)\}$ and $\text{prox}_{\lambda||x||_2}\{f(x)\}$.

Solution:

By definition

$$\text{prox}_{\lambda h(x)}\{f(x)\} = \text{argmin}_v\{h(v) + \frac{1}{2\lambda}||x - v||_2^2\}$$

Let $h(x) = ||x||_1$

$$\text{prox}_{\lambda||x||_1}\{f(x)\} = \text{argmin}_v\{||v||_1 + \frac{1}{2\lambda}||x - v||_2^2\}$$

With insight we anticipate that the optimum of mixed $\ell_1 - \ell_2$ norms is given by the soft-threshold or shrinkage operator.

For this problem we can use an extension of the optimality conditions to sub-differentiable functions.

$$0 \in \partial_v F = \partial_v[||v||_1 + \frac{1}{2\lambda}||v - x||_2^2]$$
$$0 \in \partial_v F = \partial_v||v||_1 + \frac{1}{2\lambda}\partial_v||v - x||_2^2$$
$$0 \in \partial_v F = \partial_v||v||_1 + \frac{1}{2\lambda}\nabla||v - x||_2^2$$
$$0 \in \lambda\partial_v||v||_1 + v - x$$

Now we consider the sub-differential for ℓ_1 component wise.

$$\partial_v ||v||_1 = \begin{cases} \text{sign}(v_i) & \text{for } v_i \neq 0 \\ [-1, 1] & \text{for } v_i = 0 \end{cases}$$

Analyzing both cases we have the following.

$$\begin{cases} 0 = v_i^* - x + \lambda \text{sign}(v_i^*) & v_i \neq 0 \\ 0 \in x + \lambda[-1, 1] & v_i = 0 \end{cases}$$

Solving for the minimizer v^* in terms of x.

$$\begin{cases} v_i^* = x - \lambda \text{sign}(v_i^*) & v_i \neq 0 \\ x \in \lambda[-1, 1] & v_i = 0 \end{cases}$$

From the first condition we see that if $v_i^* \leq 0$ then $x \leq 0$ (notice that $\lambda > 0$).

$$0 > v^* = x + \lambda$$

Similarly for $v^* > 0$ the $x > 0$.

$$0 < v^* = x - \lambda$$

Now using the fact that x and v^* have similar signs we may write the solution for v^* exclusively in terms of x.

$$v_i = \begin{cases} 0 & x \in [-\lambda, \lambda] \\ x - \lambda \text{sign}(x) & \text{otherwise} \end{cases}$$

This is exactly the shrinkage operator we anticipated to find.

Now consider $\text{prox}_{\lambda||x||_2}\{x\}$

Again by definition

$$\text{prox}_{\lambda h(x)}\{f(x)\} = \text{argmin}_v\{h(v) + \frac{1}{2\lambda}||x - v||_2^2\}$$

Let $h(x) = ||x||_2$

$$\text{prox}_{\lambda||x||_2}\{f(x)\} = \text{argmin}_v\{||v||_2 + \frac{1}{2\lambda}||x - v||_2^2\}$$

In this instance the function is differentiable everywhere.

$$
\begin{aligned}
0 &= \nabla[||v||_2 + \frac{1}{2\lambda}||v - x||_2^2] \\
0 &= \nabla||v||_2 + \frac{1}{2\lambda}\nabla||v - x||_2^2 \\
0 &= \frac{1}{2}v + \frac{1}{\lambda}v - x \\
(\frac{1}{2} + \lambda^{-1})v &= \lambda^{-1}x \\
v &= \frac{2}{2+\lambda}x
\end{aligned}
$$

Thus the optimal value is given by $v^* = \frac{2}{2+\lambda}x$.

2. For $\lambda = 0.001$, numerically solve the problem $\min_x F(x)$ using sub-gradient method, proximal gradient method, accelerated proximal gradient method with heavy-ball momentum and Nesterov's acceleration. Plot $F(x^k) - F(x^*)$ over the iteration k for each method, where x^* is in the code that used to generate the training data.

Solution:

For the optimization problem $F(x) = f(x) + \lambda R(x)$ we implement the following iterative scheme for subgradient descent.

$$x_{t+1} = x_t - \eta_t \partial F$$

In the case that $R(x) = ||x||_2^2$, the function is everywhere differentiable and we have the following

$$\begin{aligned}
x_{t+1} &= x_t - \eta_t \partial F(x_t) \\
x_{t+1} &= x_t - \eta_t \nabla F(x_t) \\
x_{t+1} &= x_t - \eta_t \left[\nabla \frac{1}{2n} \sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T x_t)) + \lambda \nabla ||x_t||_2^2 \right] \\
x_{t+1} &= x_t - \eta_t \left[\frac{1}{2n} \sum_{i=1}^{n} \frac{-b_i a_i \exp(-b_i a_i^T x_t)}{1 + \exp(-b_i a_i^T x_t)} + 2\lambda x_t \right]
\end{aligned}$$

In the case that $R(x) = ||x||_1^2$, the function is not differentiable everywhere thus we break the subgradient into cases component wise.

$$\begin{aligned}
x_{t+1} &= x_t - \eta_t \partial F(x_t) \\
x_{t+1} &= x_t - \eta_t \left[\nabla \frac{1}{2n} \sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T x_t)) + \lambda \partial ||x_t||_1 \right] \\
x_{t+1} &= x_t - \eta_t \left[\frac{1}{2n} \sum_{i=1}^{n} \frac{-b_i a_i \exp(-b_i a_i^T x_t)}{1 + \exp(-b_i a_i^T x_t)} + \lambda \text{sign}(x_i) \right]
\end{aligned}$$

As we showed in class we expect that the subgradient with stepsize $\eta_t \approx \frac{1}{\sqrt{t}}$ has a sublinear convergence rate of $\mathcal{O}(\frac{1}{\sqrt{t}})$. This is what we observe in the following plot.

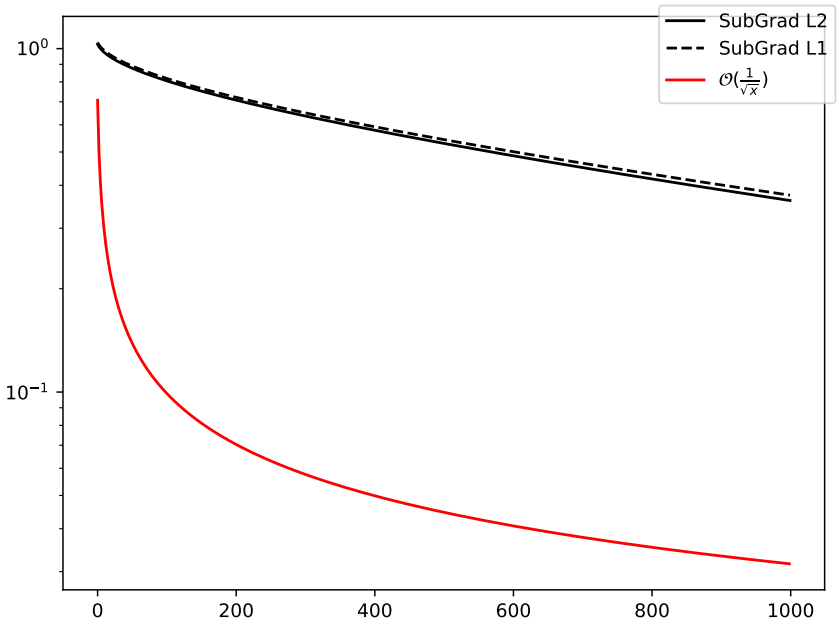

Figure 1: Order of Subgradient Method

For the optimization problem $F(x) = f(x) + \lambda h(x)$ we implement the following iterative scheme for proximal gradient descent.

$$x_{t+1} = \text{prox}_{\lambda h(x)}\{x_t - \eta_t \partial f\}$$

In the case that $h(x) = ||x||_2$, the proximal function is everywhere differentiable. For simplicity let $h(x) = ||x||_2^2$ and $\lambda = 2$.

$$\begin{aligned}
x_{t+1} &= \text{prox}_{\lambda h(x)}\{x_t - \eta_t \partial f\} \\
x_{t+1} &= \text{argmin}_v\{||v - x_t + \eta_t \partial f||_2^2 + \frac{\lambda}{2}||v||_2^2 \\
0 &= v^* - x_t + \eta_t \partial f + \lambda v^* \\
v* &= \frac{x_t - \eta_t \partial f}{1 + \lambda} \\
x_{t+1} = v^* &= \frac{x_t - \eta_t \partial f}{1 + \lambda}
\end{aligned}$$

In the case that $h(x) = ||x||_1$, the proximal function is not differentiable everywhere. Therefore we must consider the function in cases.

$$x_{t+1} = \text{prox}_{\lambda h(x)}\{x_t - \eta_t \partial f\}$$

$$x_{t+1} = \text{argmin}_v\{||v - x_t + \eta_t \partial f||_2^2 + \lambda||v||_1$$

In part 1 we showed that the solution to this problem is the soft-threshold algorithm

$$x = \begin{cases} 0 & x_i - \eta_t \nabla f(x)_i \in [-\lambda, \lambda] \\ x_i - \eta \nabla f(x)_i - \lambda \text{sign}(x_i - \eta_t \nabla f(x)_i) & \text{otherwise} \end{cases}$$

If $f(x) + h(x)$ is L-smooth and $\eta_t \equiv 1/L$ then we can anticipate that the method is order $\mathcal{O}(\frac{1}{t})$. This is not the case for $h(x) = ||x||_1$, however it is true for $h(x) = ||x||_2^2$. Additionally (as we will see later) the value of lambda will force the $||x||_1$ regularization term to dominate after a certain number of iterations and the value of x^* will become zero.

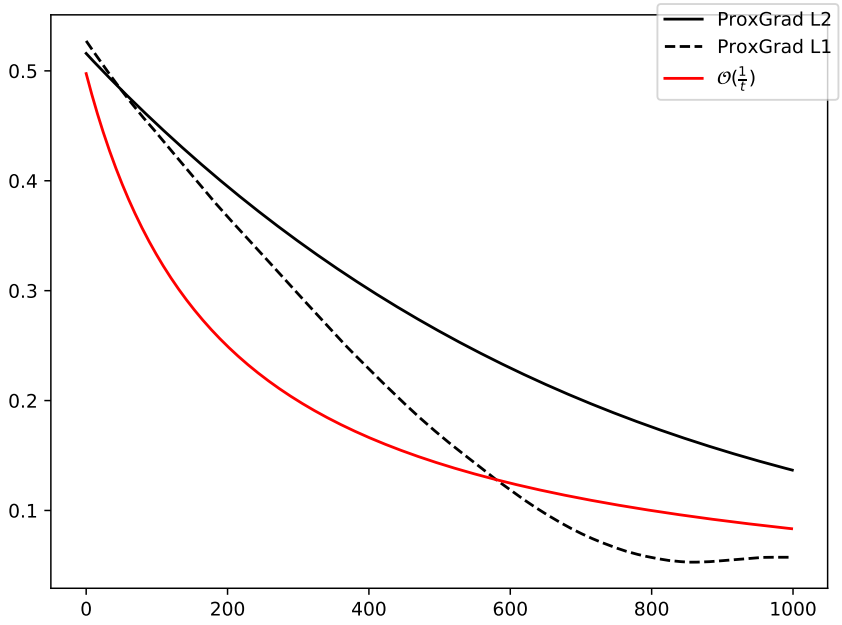

Figure 2: Order of Proximal Gradient Method

For the optimization problem $F(x) = f(x) + \lambda h(x)$ we implement the following iterative scheme for momentum proximal gradient descent.

$$x_{t+1} = \text{prox}_{\lambda h(x)}\{x_t - \eta_t \partial f + \theta(x_t - x_{t-1})\}$$

In the case that $h(x) = ||x||_2$, the proximal function is everywhere differentiable. Similar to the case above we find the following iterative scheme.

$$x_{t+1} = \frac{x_t - \eta_t \partial f + \theta(x_t - x_{t-1})}{1 + \lambda}$$

Similarly if $h(x) = ||x||_1$, the proximal function has similar analysis to the non-accelerated case.

$$x^{t+1} = \begin{cases} 0 & x_i^t - \eta_t \nabla f(x^t)_i + \theta_t(x_i^t - x_i^{t-1}) \in [-\lambda, \lambda] \\ x_i^t - \eta \nabla f(x^t)_i - & \\ \quad -\lambda \text{sign}(x_i^t - \eta_t \nabla f(x^t)_i + \theta_t(x_i^t - x_i^{t-1})) & \text{otherwise} \end{cases}$$

If $f(x) + h(x)$ is L-smooth and $\eta_t \equiv 1/L$ then we can anticipate that the method is order $\mathcal{O}(\frac{1}{(t+1)^2})$.

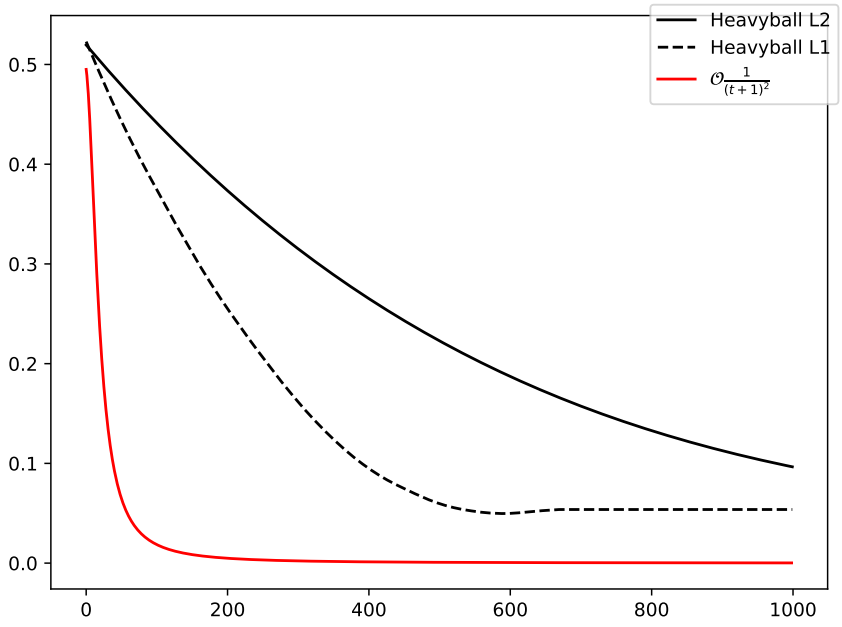

Figure 3: Order of Momentum Accelerated Proximal Gradient Method

---

Finally for the Nesterov accelerated proximal algorithm we implement the following iterative scheme.

$$y_t = x_t + \frac{t}{t+3}(x_t - x_{t-1})$$
$$x_{t+1} = \text{prox}_{\lambda h(x)}\{y_t - \eta_t \partial f(y_t)\}$$

Although this is not a descent algorithm it can be shown that it maintains $\mathcal{O}(\frac{1}{(t+1)^2})$ convergence for $\eta_t \equiv 1/L$ if f is L-smooth.

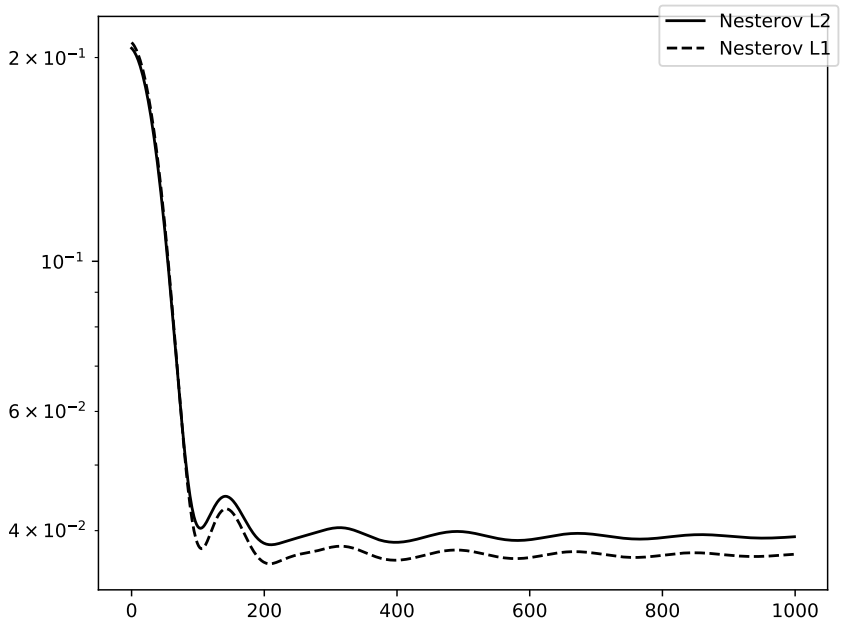

Figure 4: Nesterov Oscillations

It is important to note here that the minimizer x^* suggested to use in this model is a minimizer of the MLE function $f(x)$ without regularization. Therefore for any nonzero value of λ we can anticipate that the true solution is different than x^*. Thus explaining why if we compare $F(x) - F(x^*)$ we see the oscillating behavior of the Nesterov acceleration but we do not see the continued convergence pattern.

3. Test different λ, e.g. $0.005, 0.01, 0.05, 0.1$ and see how x^k changes after you run enough number of iterations.

Solution:

Consider the coefficients of the logistic regression function. Because of the regularization terms the values of the solution are penalized for being particularly large. For instance the first term of the coefficients is 1. However, when regularizing based on the norm of x values tend towards zero.

As λ increases the values of x tend towards zero. If the regularization term is the ℓ_2 norm then all values tend towards zero. If the regularization term is the ℓ_1 norm then all values tend towards zero.

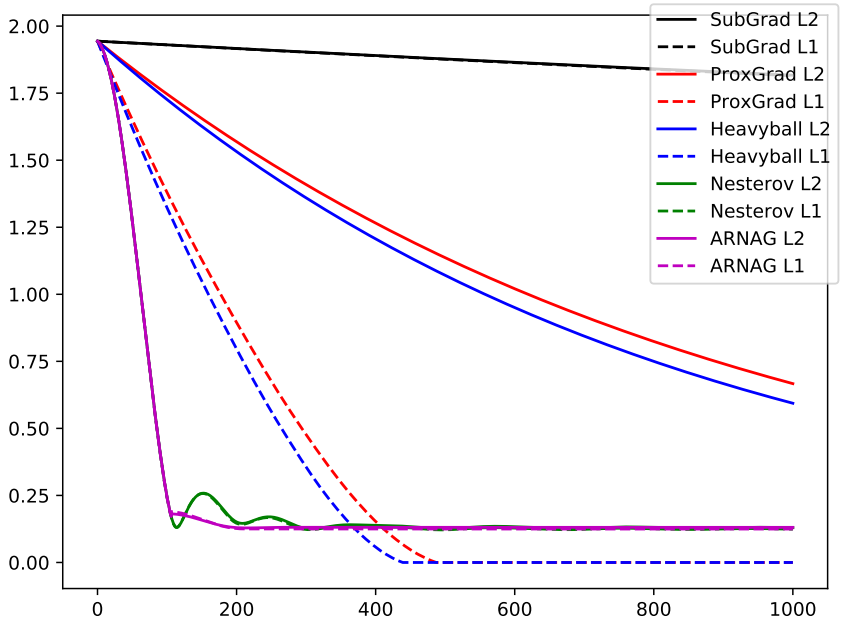

Figure 5: L2 Norm of x^* using $R(x) = ||x||_2, \lambda = 0.001$

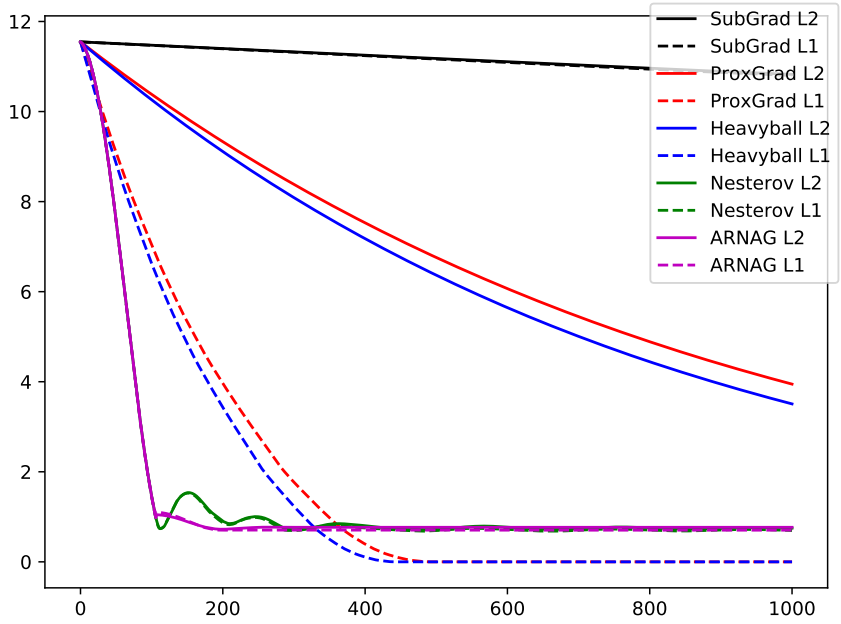

Figure 6: L1 Norm of x^* using $R(x) = ||x||_1, \lambda = 0.001$

In the above figures the norm of x^* approaches zero much more slowly than in the following figures. This corresponds to larger λ penalizing the value of x^* more heavily. For a full set of figures please see the appendix.

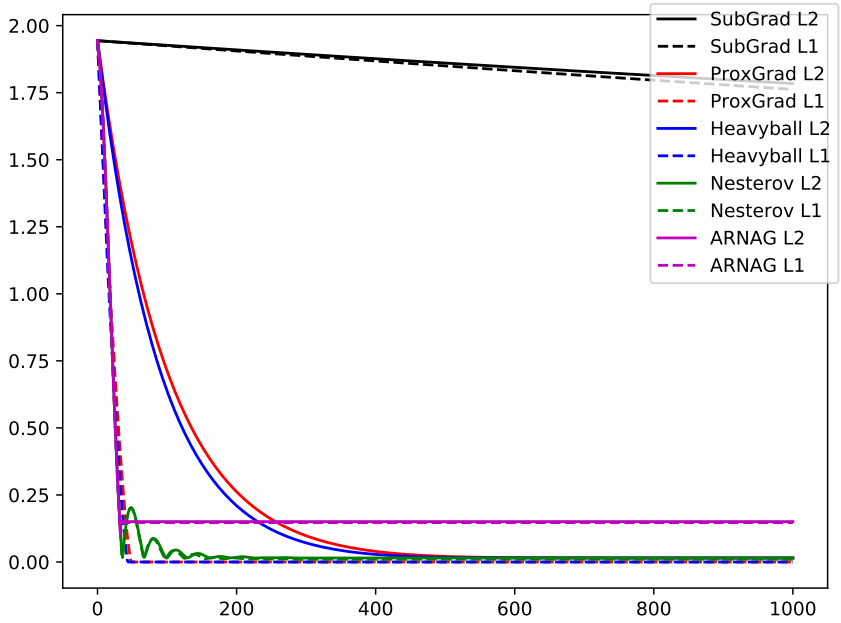

Figure 7: L2 Norm of x^* using $R(x) = ||x||_2, \lambda = 0.01$

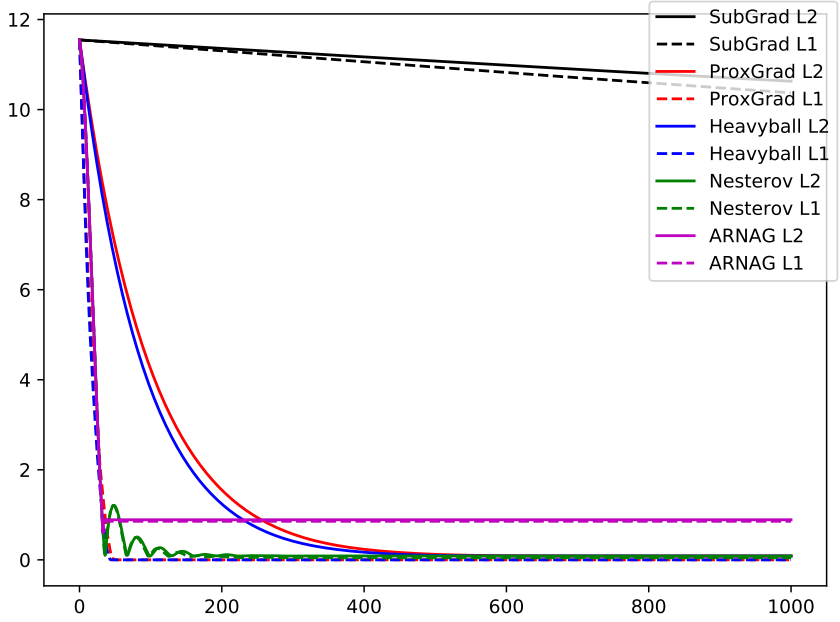

Figure 8: L1 Norm of x^* using $R(x) = ||x||_1, \lambda = 0.01$

4. Can you propose any approach to further accelerate the training process?

Solution:

One way to improve the methods is to use adaptive restart in Nesterov accelerated proximal gradient. This idea follows that from "Sharpness, Restart and Acceleration" by Roulet and d'Aspremont (NeurIPs 2017). Because the Nesterov accelerated gradient is not a descent algorithm, it may suffer when $\frac{t}{t+3} \sim 1$. Therefore, if $\nabla f(x_t - x_{t-1}) < 0$ we reset the value of t to zero. This is shown to have linear convergence if f is L-smooth.

Although the problem described above hides the true performance of the methods, in the figure below you can see that the adaptive restart performs better than standard Nestorov. Additionally, it does not have oscillations as Nestorov does.

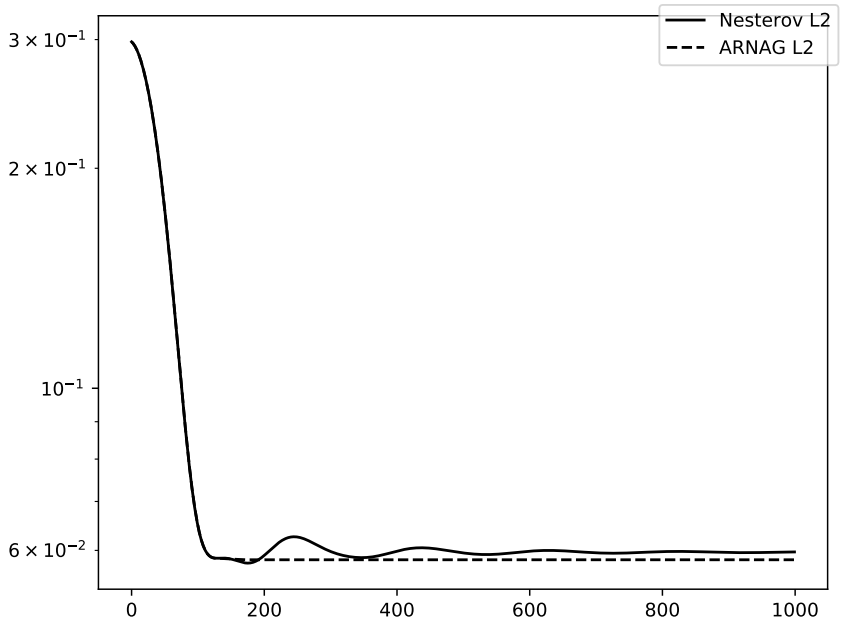

Figure 9: Adaptive Restart of Nestorov Accelerated Gradient

Comparing all of these methods together we have the following figure.

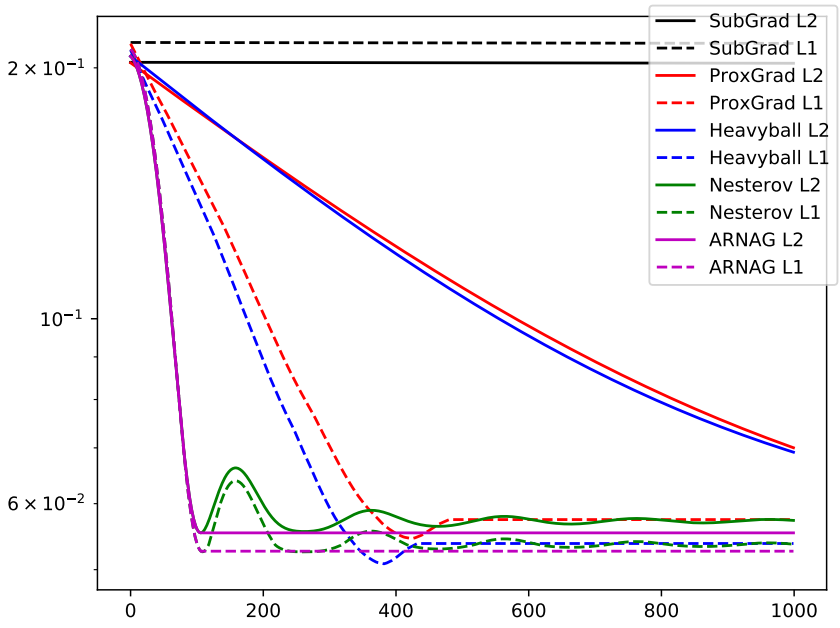

Figure 10: Comparison of all methods

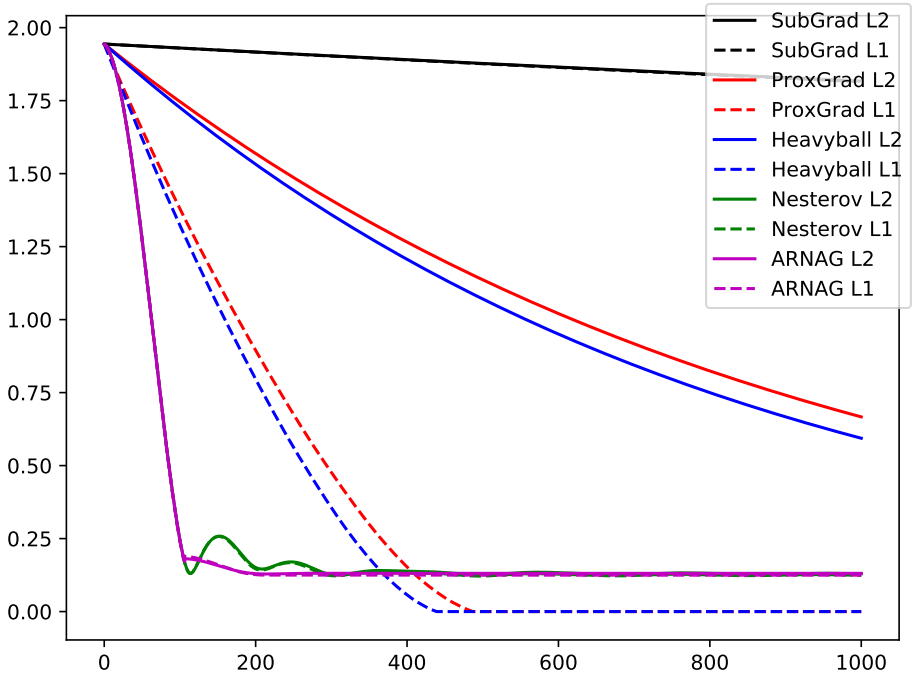

Figure 11: L2 Norm of x^* using $R(x) = ||x||_2, \lambda = 0.001$

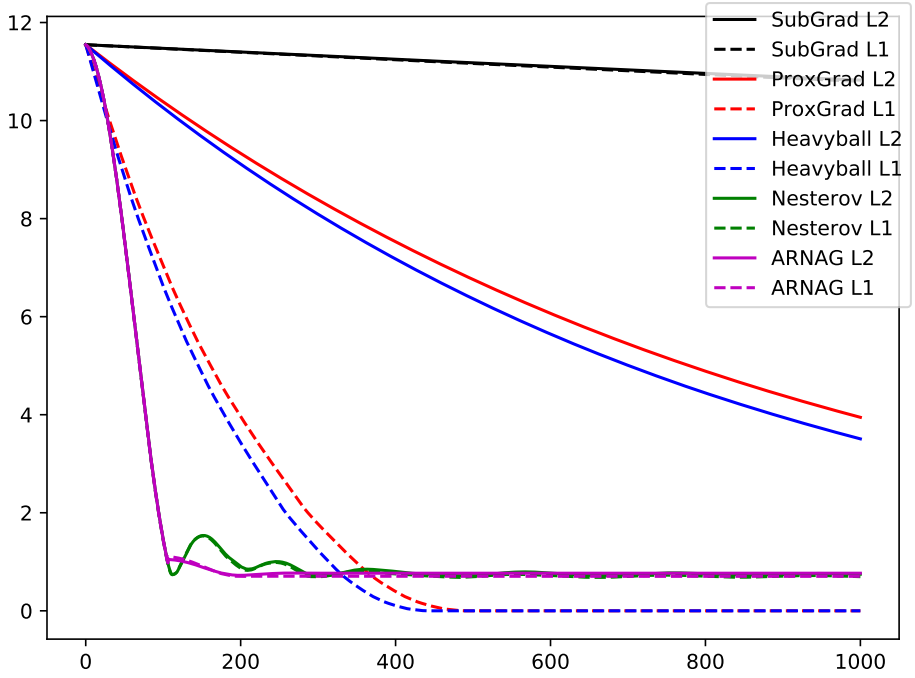

Figure 12: L1 Norm of x^* using $R(x) = ||x||_1, \lambda = 0.001$

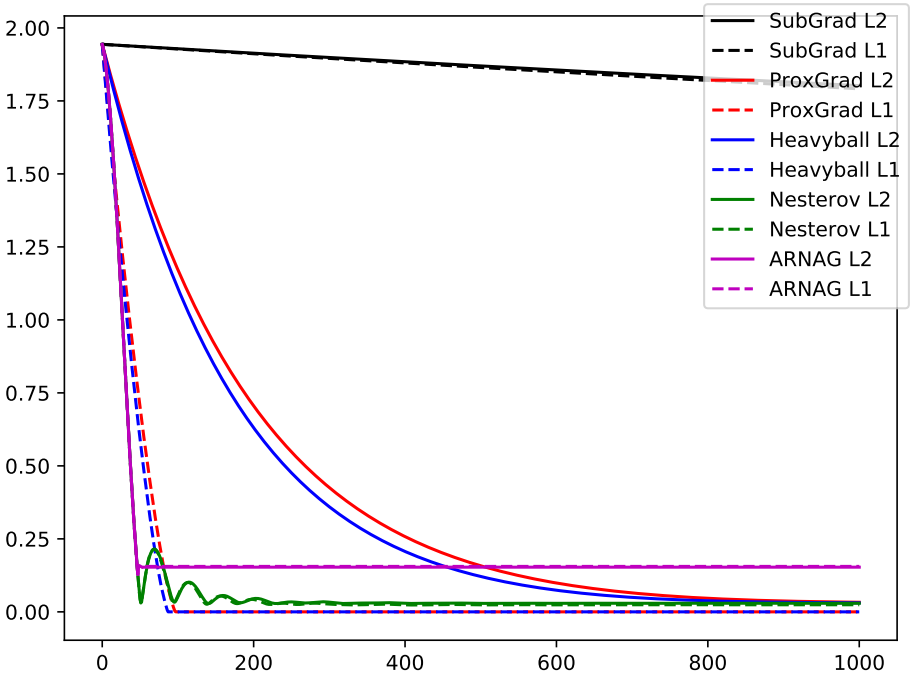

Figure 13: L2 Norm of x^* using $R(x) = ||x||_2, \lambda = 0.005$

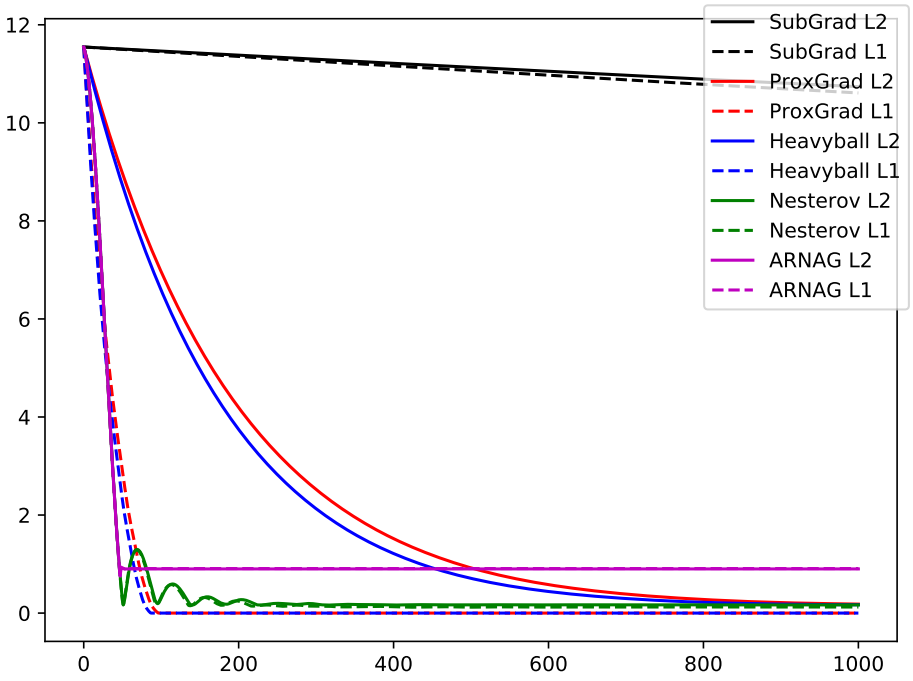

Figure 14: L1 Norm of x^* using $R(x) = ||x||_1, \lambda = 0.005$

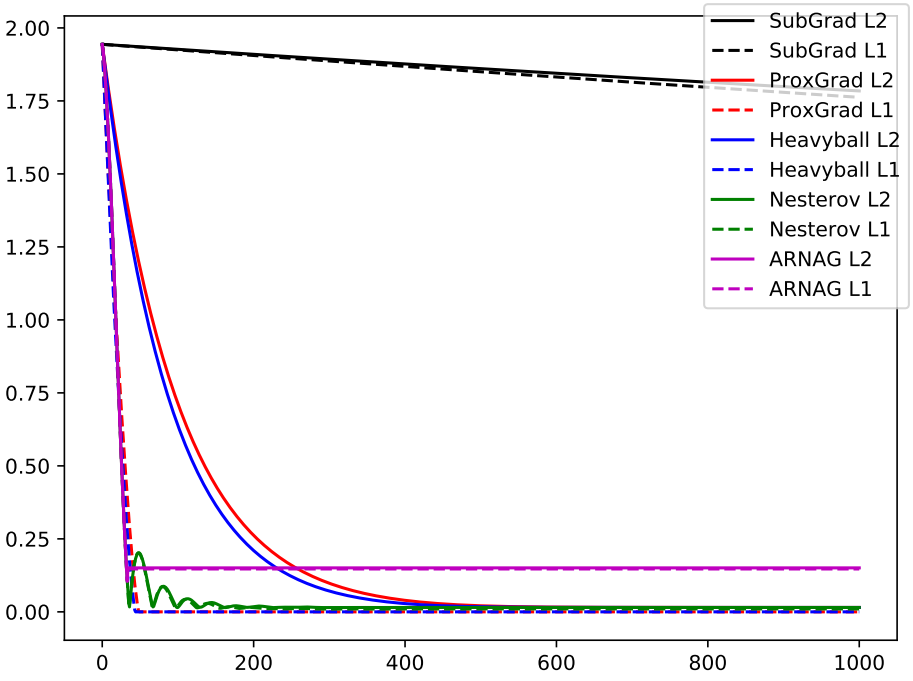

Figure 15: L2 Norm of x^* using $R(x) = ||x||_2, \lambda = 0.01$

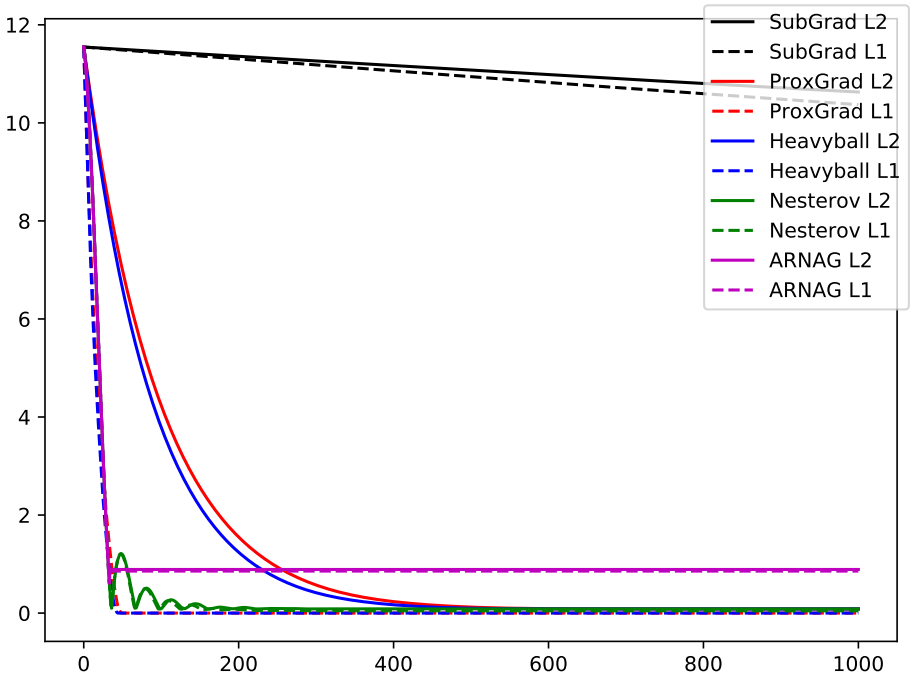

Figure 16: L1 Norm of x^* using $R(x) = ||x||_1, \lambda = 0.01$

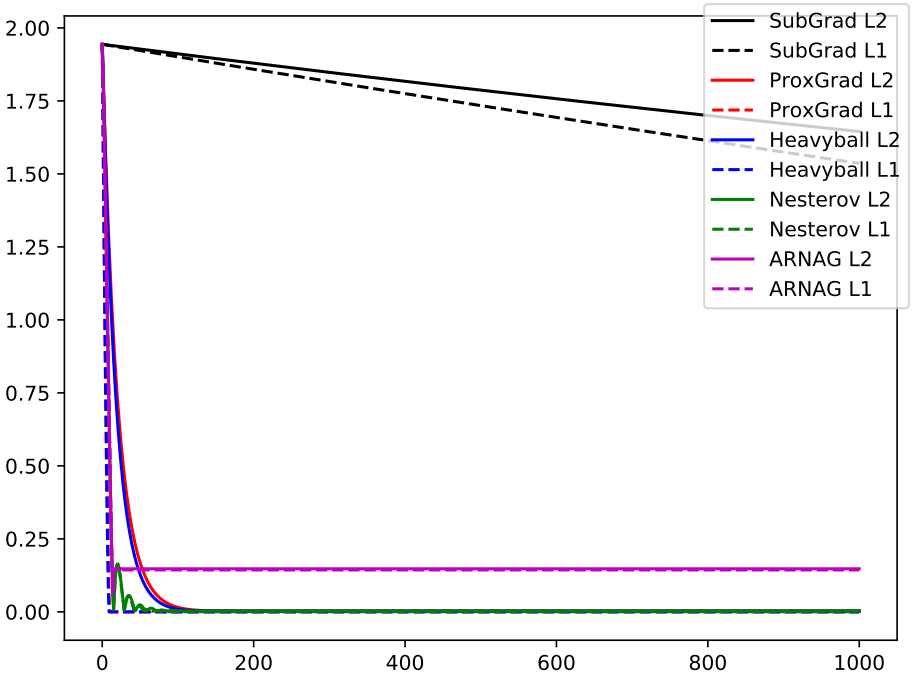

Figure 17: L2 Norm of x^* using $R(x) = ||x||_2, \lambda = 0.05$

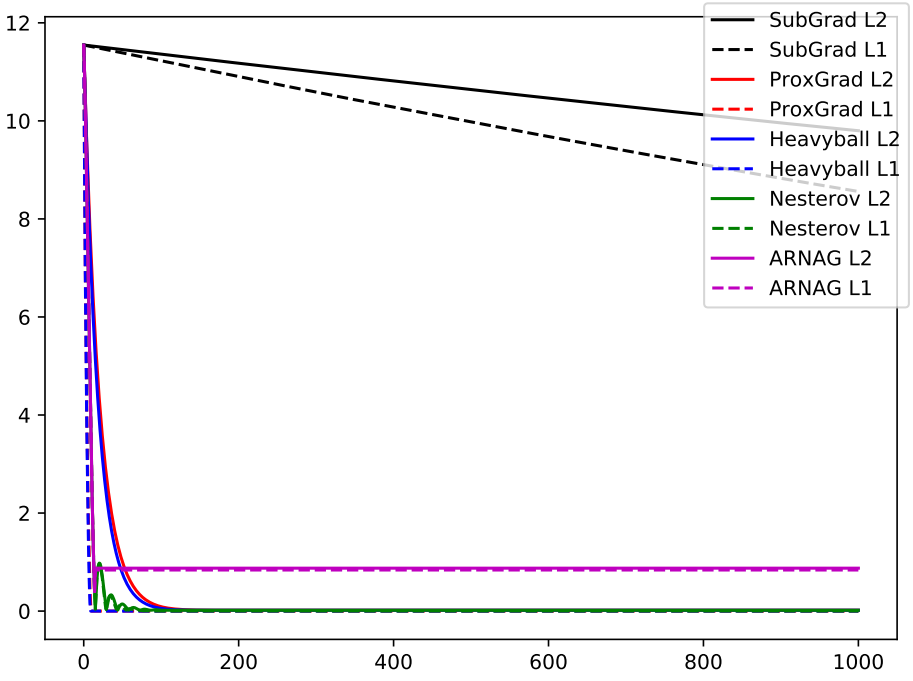

Figure 18: L1 Norm of x^* using $R(x) = ||x||_1, \lambda = 0.05$

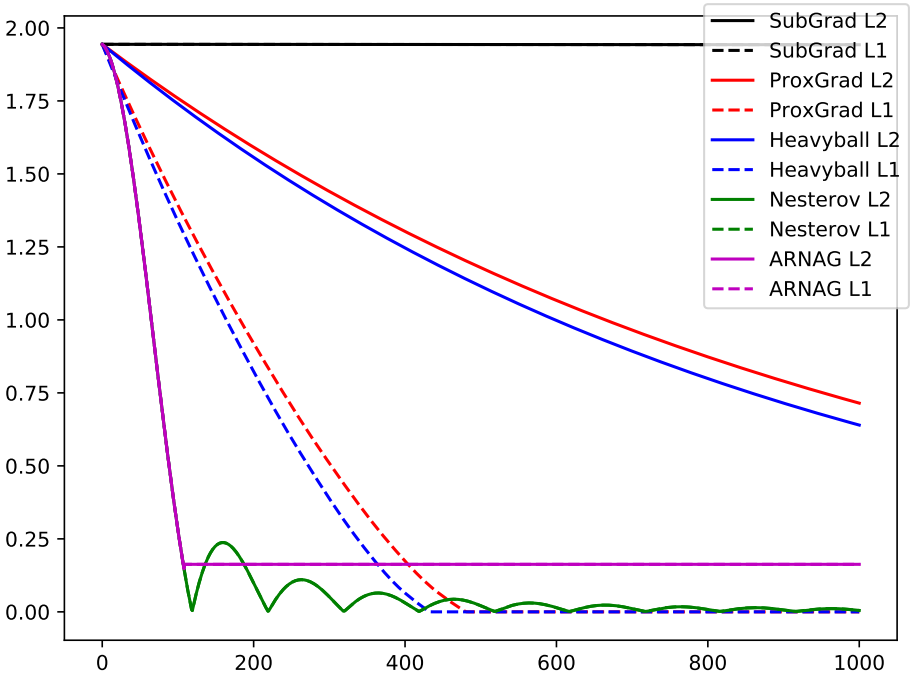

Figure 19: L2 Norm of x^* using $R(x) = ||x||_2, \lambda = 0.1$

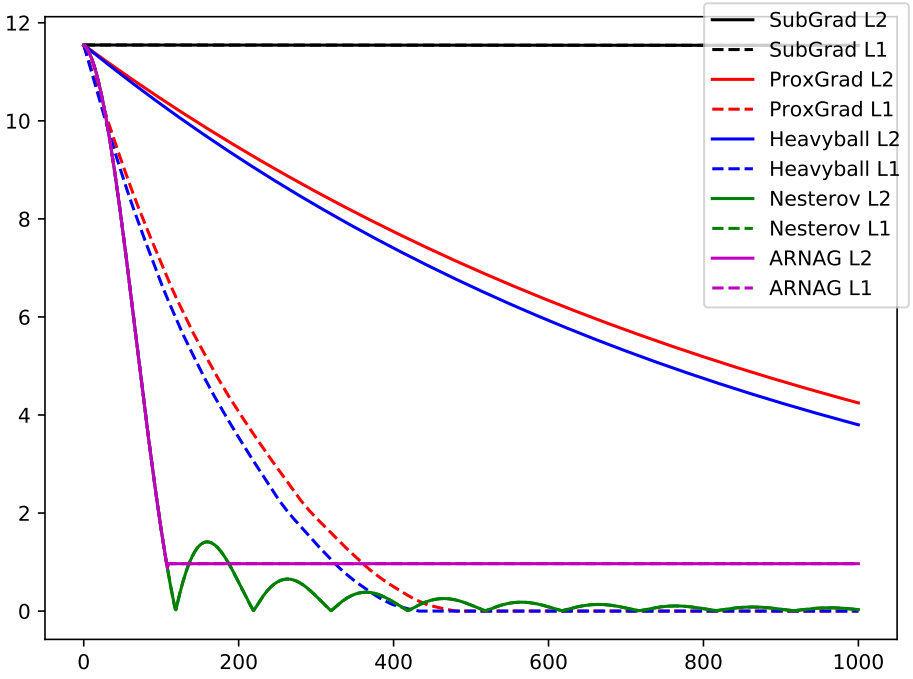

Figure 20: L1 Norm of x^* using $R(x) = ||x||_1, \lambda = 0.1$