HOOML2022: Order up! The Benefits of Higher-Order Optimization in Machine Learning 1–36

Using quadratic equations for overparametrized models

Shuang Li SHUANGLI@MATH.UCLA.EDU
UCLA, Los Angeles, USA
William J. Swartworth WSWARTWORTH@GMAIL.COM
UCLA, Los Angeles, USA
Martin Takáč TAKAC.MT@GMAIL.COM
MBZUAI, Abu Dhabi, UAE
Deanna Needell DEANNA@MATH.UCLA.EDU
UCLA, Los Angeles, USA
Robert M. Gower RGOWER@FLATIRONINSTITUTE.ORG
Center for Computational Mathematics, Flatiron Institute, Simons Foundation, New York, NY, USA

Abstract

Recently the `SP` (Stochastic Polyak step size) method has emerged as a competitive adaptive method for setting the step sizes of SGD. `SP` can be interpreted as a method specialized to interpolated models, since it solves the *interpolation equations*. `SP` solves these equation by using local linearizations of the model. We take a step further and develop a method for solving the interpolation equations that uses the local second-order approximation of the model. Our resulting method `SP2` uses Hessian-vector products to speed-up the convergence of `SP`. Furthermore, and rather uniquely among second-order methods, the design of SP2 in no way relies on positive definite Hessian matrices or convexity of the objective function. We show `SP2` is very competitive on matrix completion, non-convex test problems and logistic regression. We also provide a convergence theory on sums-of-quadratics.

1. Introduction

Consider the problem

$$w^* \in \operatorname{argmin}_{w\in\mathbb{R}^d} \left\{ f(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w) \right\}, \tag{1}$$

where f is twice continuously differentiable, and the set of minimizers is nonempty. Let the optimal value of (1) be $f^* \in \mathbb{R}$, and w^0 be a given initial point. Here each $f_i(w)$ is the loss of a model parametrized in $w \in \mathbb{R}^d$ over an i-th data point. Our discussion, and forth coming results, also hold for a loss given as an expectation $f(w) = \mathbf{E}_{\xi\sim\mathcal{D}}[f_\xi(w)]$, where $\xi \sim \mathcal{D}$ is the data generating process and $f_\xi(w)$ the loss over this sampled data point. But for simplicity we use the $f_i(w)$ notation.

Contrary to classic statistical modeling, there is now a growing trend of using over-parametrized models that are able to interpolate the data Ma et al. (2018); that is, models for which the loss is minimized over every data point as described in the following assumption.

Assumption 1 *We say that the interpolation condition holds when the loss is nonnegative,* $f_i(w) \geq 0$, *and*

$$\exists\ w^* \in \mathbb{R}^d \quad \textit{such that} \quad f(w^*) = 0. \tag{2}$$

Consequently, $f_i(w^*) = 0$ for $i = 1, \ldots, n$. Overparameterized deep neural networks are the most notorious example of models that satisfy Assumption 1. Indeed, with sufficiently more parameters than data points, we are able to simultaneously minimize the loss over all data points.

If we admit that our model can interpolate the data, then we have that our optimization problem (1) is equivalent to solving the system of nonlinear equations

$$f_i(w) \; = \; 0, \quad \text{for } i = 1, \ldots, n. \tag{3}$$

Since we assume $f_i(w) \geq 0$ any solution to the above is a solution to our original problem.

Recently, it was shown in Berrada et al. (2020); Gower et al. (2021a) that the Stochastic Polyak step size (`SP`) method Loizou et al. (2020); Polyak (1987) directly solves the interpolation equations. Indeed, at each iteration `SP` samples a single i-th equation from (3), then projects the current iterate w^t onto the linearization of this constraint, that is

$$w^{t+1} = \mathrm{argmin}_{w \in \mathbb{R}^d} \left\| w - w^t \right\|^2 \qquad \text{s.t. } \; f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle = 0. \tag{4}$$

The closed form solution to (4) is given by

$$w^{t+1} = w^t - \tfrac{f_i(w^t)}{\left\| \nabla f_i(w^t) \right\|^2} \nabla f_i(w^t). \tag{5}$$

Here we take one step further, and instead of projecting onto the linearization of $f_i(w)$ we use the local quadratic expansion. That is, as a proxy of setting $f_i(w) = 0$ we set the quadratic expansion of $f_i(w)$ around w^t to zero

$$f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle + \tfrac{1}{2} \left\langle \nabla^2 f_i(w^t)(w - w^t), w - w^t \right\rangle = 0. \tag{6}$$

The above quadratic constraint could have infinite solutions, a unique solution or no solution at all[1]. Indeed, for example if $\nabla^2 f_i(w^t)$ is positive definite, there may exist no solution, which occurs when f_i is convex, and is the most studied setting for second order methods. But if the loss is positive f_i and the Hessian has at least one negative eigenvalue, then (6) always has a solution.

If (6) has solutions, then analogously to the `SP` method, we can choose one using a projection step [2]

$$\begin{aligned} w^{t+1} \; \in \; & \mathrm{argmin}_{w \in \mathbb{R}^d} \tfrac{1}{2} \left\| w - w^t \right\|^2 \\ & \text{s.t. } \; f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle + \tfrac{1}{2} \left\langle \nabla^2 f_i(w^t)(w - w^t), w - w^t \right\rangle = 0. \end{aligned} \tag{7}$$

We refer to (7) as the `SP2` method. Using a quadratic expansion has several advantages. First, quadratic expansions are more accurate than linearizations, which will allow us to take larger steps. Furthermore, using the quadratic expansion will lead to convergence rates which are *independent* on how well conditioned the Hessian matrices are, as we show later in Proposition 5.

1. Or even two solutions in the 1d case.
2. Note that there could be more than one solution to this projection and we can choose one either with least norm or arbitrarily.

Our SP2 method occupies a unique position in the literature of stochastic second order method since it is incremental and in no way relies on convexity or positive semi-definite Hessian matrices. Indeed, as we will show in our non-convex experiments in 3.1 and matrix completition E, the SP2 excels at minimizing non-convex problems that satisfy interpolation. In contrast, Newton based methods often converge to stationary points other than the global minima.

The rest of this paper is organized as follows. We introduce some related work in Section A. We present the proposed SP2 methods in Section 2 and corresponding convergence analysis in Section B. In Section C, we relax the interpolation condition and develop a variety of quadratic methods to solve the slack version of this problem. We test the proposed methods with a series of experiments in Section 3. Finally, we conclude our work and discuss future directions in Section 4.

2. The SP2 Method

Next we give a closed form solution to (7) for GLMs. We then provide an approximate solution to (7) for more general models.

2.1. SP2 - Generalized Linear Models

Consider when f_i is the loss over a linear model with

$$f_i(w) = \phi_i(x_i^\top w - y_i), \tag{8}$$

where $\phi_i : \mathbb{R} \to \mathbb{R}$ is a loss function, and $(x_i, y_i) \in \mathbb{R}^{d+1}$ is an input-output pair. Consequently

$$\nabla f_i(w) = \phi_i'(x_i^\top w - y_i)x_i \coloneqq a_i x_i, \qquad \nabla^2 f_i(w) = \phi_i''(x_i^\top w - y_i)x_i x_i^\top \coloneqq h_i x_i x_i^\top. \tag{9}$$

The quadratic constraint problem (7) can be solved exactly for GLMs (8) as we show next.

Lemma 2 *(SP2) Assume $f_i(w)$ is the loss of a generalized linear model (8) and is non-negative. Let $f_i = f_i(w^t)$ for short. Let $a_i \coloneqq \phi_i'(x_i^\top w - y_i)$ and $h_i \coloneqq \phi_i''(x_i^\top w - y_i)$. If*

$$a_i^2 - 2h_i f_i \geq 0 \tag{10}$$

then the optimal solution of (7) is as follows

$$w^{t+1} = w^t - \frac{a_i}{h_i}\left(1 - \frac{1}{|a_i|}\sqrt{a_i^2 - 2h_i f_i}\right)\frac{x_i}{\|x_i\|^2}. \tag{11}$$

Alternatively if (10) does not hold, since $f_i \geq 0$ we have necessarily that $h_i > 0$, and consequently a Newton step will give the minima of the local quadratic, that is

$$w^{t+1} = w^t - \frac{a_i}{h_i}\frac{x_i}{\|x_i\|^2}. \tag{12}$$

The proof to the above lemma, and all subsequent missing proofs can be found in the appendix. Lemma (2) establishes a condition (10) under which we should not take a full Newton step. Interestingly, this condition (10) holds when the square root of the loss function has negative curvature, as we show in the next lemma.

Lemma 3 *Let ϕ be a non-negative function which is twice differentiable at all t with $\phi(t) \neq 0$. The condition (10), in other words*

$$\phi'(t)^2 \geq 2\phi(t)\phi''(t)$$

holds when $\sqrt{\phi(t)}$ is concave away from its roots, i.e. when $\frac{d^2}{dt^2}\sqrt{\phi(t)} \leq 0$ for all t with $\phi(t) \neq 0$.

Examples of loss functions include $\phi(t) = \tanh^2(t)$ and $\phi(t) = t^p$ with $0 \leq p \leq 2$.

In conclusion to this section, SP2 has a closed form solution for GLMs, and this closed form solution includes many non-convex loss functions.

2.2. SP2$^+$ - Linearizing and Projecting

In general, there is no closed form solution to (7). Indeed, there may not even exist a solution. Inspired by the fact that computing a Hessian-vector product can be done with a single backpropagation at the same cost as computing a gradient Christianson (1992), we will make use of the cheap Hessian-vector product to derive an approximate solution to (7).

Instead of solving (7) exactly, here we propose to take two steps towards solving (7) by projecting onto the linearized constraints. To describe this method let

$$q(w) := f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle + \tfrac{1}{2}\left\langle \nabla^2 f_i(w^t)(w - w^t), w - w^t \right\rangle. \tag{13}$$

In the first step we linearize the quadratic constraint (13) around w^t and project onto this linearization:

$$w^{t+1/2} = \mathrm{argmin}_{w \in \mathbb{R}^d} \tfrac{1}{2}\left\|w - w^t\right\|^2 \quad \text{s.t. } f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle = 0. \tag{14}$$

The closed form update of this first step is given by

$$w^{t+1/2} \;=\; w^t - \frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t), \tag{15}$$

which is a Stochastic Polyak step (5). For the second step, we once again linearize the quadratic constraint (13), but this time around the point $w^{t+1/2}$ and set this linearization to zero, that is

$$w^{t+1} = \mathrm{argmin}_{w \in \mathbb{R}^d} \tfrac{1}{2}\left\|w - w^{t+1/2}\right\|^2 \;\text{ s.t. } q(w^{t+1/2}) + \left\langle \nabla q(w^{t+1/2}), w - w^{t+1/2} \right\rangle = 0. \tag{16}$$

The closed form update of this second step is given by

$$w^{t+1} = w^{t+1/2} - \frac{q(w^{t+1/2})}{\left\|\nabla q(w^{t+1/2}))\right\|^2}\nabla q(w^{t+1/2}). \tag{17}$$

We refer to the resulting proposed method as the SP2$^+$ method, summarized in the following.

Lemma 4 *(SP2$^+$) Let $g_t \equiv \nabla f_i(w^t)$ and $\mathbf{H}_t \equiv \nabla^2 f_i(w^t)$. Update (15) and (17) is given by*

$$w^{t+1} = w^t - \frac{f_i(w^t)}{\|g_t\|^2}g_t - \frac{1}{2}\frac{f_i(w^t)^2}{\|g_t\|^4}\frac{\langle \mathbf{H}_t g_t, g_t\rangle}{\|v^{t+1}\|^2}v^{t+1}, \text{ where } v^{t+1} = \left(\mathbf{I} - \mathbf{H}_t\frac{f_i(w^t)}{\|g_t\|^2}\right)g_t, \tag{18}$$

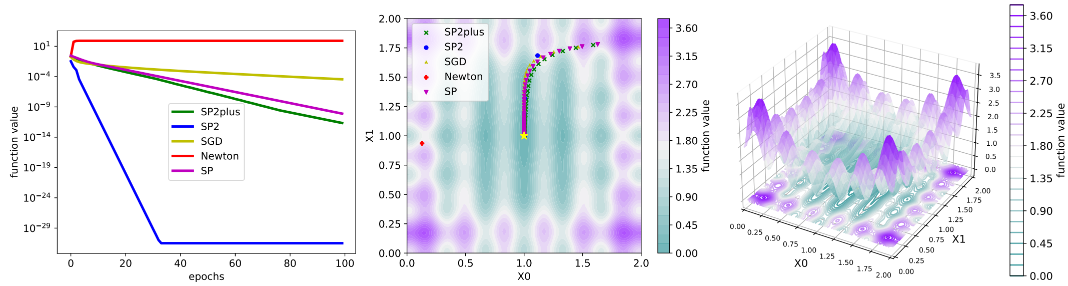


Figure 1: The Levy N. 13 function where Left: we plot $f(x)$ across epochs Middle: level set plot, Right: Surface plot. SP2 is in blue , $\texttt{SP2}^+$ is in green , SGD is in yellow and Newton is in red .

In (18) we can see that $\texttt{SP2}^+$ applies a second order correction to the SP step.

$\texttt{SP2}^+$ is equivalent to two steps of a Newton Raphson method applied to finding a root of $q(w)$. If we apply multiple steps of the Newton Raphson method, as opposed to two, the resulting method converges to the root of q, see Theorem 11 in the appendix. Theorem 11 shows that this multi-step version of $\texttt{SP2}^+$ converges when q belongs to a large class of non-convex functions known as the star-convex functions. Star-convexity, which is a generalization of convexity, includes several non-convex loss functions Hinder et al. (2019).

3. Experiments

3.1. Non-convex problems

To emphasize how our new SP2 methods can handle non-convexity, we have tested SP2 (7), $\texttt{SP2}^+$ (18) and SP (5) on the non-convex problems PermDβ^+, Rastrigin and Levy N. 13, and Rosenbrock (Jamil and Yang (2013)[3]), see Figure 1 in the main text, and Figures 2, 3, and 4 in the appendix. All of these functions are sums-of-terms of the format (1) and satisfy the interpolation Assumption 1. To compute the SP2 update we used ten steps of Newton's Raphson method as detailed in Section F.4. We used the same step size of 0.2 on all methods except SGD for which we had to hand tune the step size to ensure convergence. We consistently find across these non-convex problems that SP2 and $\texttt{SP2}^+$ are very competitive, with SP2 converging in under 10 epochs. Here we can clearly see that SP2 converges to a high precision solution (like most second order methods), and different than other second order methods is not attracted to local maxima or saddle points. In contrast, Newtons method converges to a local maxima on all problems excluding the Rosenbrock function in Figure 4 in the appendix. For instance, on the left of Figure 1 we can see the red dot of Newton stuck on a local maxima, despite all methods having the same initialization.

3. We used the Python Package pybenchfunction available on github Python_Benchmark_Test_Optimization_Function_Single_Objective. We also note that the PermDβ^+ implemented in this package is a modified version of the PermDβ function, as we detail in Section G.1.1.

4. Conclusion

We have proposed new incremental second order methods that exploit models that interpolate the data, or are close to interpolation. What sets our methods apart from most previous incremental second order methods is that they do not rely on convexity in their design. Quite the opposite, the `SP2` method can benefit from the Hessian having at least one negative eigenvalue. Consequently the `SP2` method excels at minimizing non-convex models that interpolate, as can be seen in Sections 3.1 and E. We also provided a convergence in Theorem 11 that shows that `SP2` and its approximation $\texttt{SP2}^+$ enjoy a significantly faster rate of convergence as compared to `SGD` for sums-of-quadratics.

References

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *Journal of Machine Learning Research*, 18(116):1–40, 2017.

Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

Albert S Berahas, Jorge Nocedal, and Martin Takáč. A multi-batch l-bfgs method for machine learning. In *The Thirtieth Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Training neural networks for and by interpolation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 799–809, 13–18 Jul 2020.

Raghu Bollapragada, Richard H Byrd, and Jorge Nocedal. Exact and inexact subsampled Newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 04 2018.

Richard H Byrd, Gillian M Chin, Will Neveitt, and Jorge Nocedal. On the use of stochastic Hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995, 2011.

Jiabin Chen, Rui Yuan, Guillaume Garrigos, and Robert M. Gower. San: Stochastic average newton algorithm for minimizing finite sums, 2021.

Bruce Christianson. Automatic Hessians by reverse accumulation. *IMA Journal of Numerical Analysis*, 12(2):135–150, 1992.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, 2006.

Yu-Hong Dai. Fast algorithms for projection on an ellipsoid. *SIAM Journal on Optimization*, 16(4):986–1006, 2006.

Murat A. Erdogdu and Andrea Montanari. Convergence rates of sub-sampled Newton methods. In *Advances in Neural Information Processing Systems 28*, pages 3052–3060. Curran Associates, Inc., 2015.

Robert M. Gower, Donald Goldfarb, and Peter Richtárik. Stochastic block BFGS: Squeezing more curvature out of data. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Robert M. Gower, Aaron Defazio, and Mike Rabbat. Stochastic polyak stepsize with a moving target. *arXiv:2106.11851*, 2021a.

Robert M. Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021b.

Robert M. Gower, Mathieu Blondel, Nidham Gazagnadou, and Fabian Pedregosa. Cutting some slack for sgd with adaptive polyak stepsizes. *arXiv:2202.12328*, 2022.

Oliver Hinder, Aaron Sidford, and Nimit Sharad Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. *arXiv preprint arXiv:1906.11985*, 2019.

Majid Jahani, Xi He, Chenxin Ma, Dheevatsa Mudigere, Aryan Mokhtari, Alejandro Ribeiro, and Martin Takac. Distributed restarting newtoncg method for large-scale empirical risk minimization. 2017.

Momin Jamil and Xin-She Yang. A literature survey of benchmark functions for global optimization problems. *CoRR*, abs/1308.4008, 2013.

Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *Advances in Neural Information Processing Systems*, 29, 2016.

Jonas Moritz Kohler and Aurélien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1895–1904, 2017.

Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtarik. Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates. *arxiv:1912.01597*, 2019.

Yang Liu and Fred Roosta. Convergence of newton-mr under inexact hessian information. *SIAM J. Optim.*, 31(1):59–90, 2021.

Nicolas Loizou, Sharan Vaswani, Issam Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. *arXiv:2002.10542*, 2020.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 3331–3340, 2018.

Aryan Mokhtari and Alejandro Ribeiro. Global convergence of online limited memory BFGS. *The Journal of Machine Learning Research*, 16:3151–3181, 2015.

Aryan Mokhtari, Mark Eisen, and Alejandro Ribeiro. Iqn: An incremental quasi-newton method with local superlinear convergence rate. *SIAM Journal on Optimization*, 28(2): 1670–1698, 2018. doi: 10.1137/17M1122943.

Philipp Moritz, Robert Nishihara, and Michael I. Jordan. A linearly-convergent stochastic L-BFGS algorithm. In *International Conference on Artificial Intelligence and Statistics*, volume 51, pages 249–258, 2016.

Jaehyun Park and Stephen Boyd. General heuristics for nonconvex quadratically constrained quadratic programming. *arXiv preprint arXiv:1703.07870*, 2017.

Mert Pilanci and Martin J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1): 205–245, 2017.

Borist Polyak. Introduction to optimization. *New York, Optimization Software*, 1987.

Zheng Qu, Peter Richtárik, Martin Takáč, and Olivier Fercoq. SDNA: Stochastic dual Newton ascent for empirical risk minimization. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Anton Rodomanov and Dmitry Kropotov. A superlinearly-convergent proximal newton-type method for the optimization of finite sums. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2597–2605. PMLR, 20–22 Jun 2016.

Farbod Roosta-Khorasani and Michael W. Mahoney. Sub-sampled newton methods. *Math. Program.*, 174(1-2):293–326, 2019.

Wilfredo Sosa and Fernanda MP Raupp. An algorithm for projecting a point onto a level set of a quadratic function. *Optimization*, pages 1–19, 2020.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, 2018.

Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu. Stochastic quasi-newton methods for nonconvex stochastic optimization, 2017.

Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A Olson, Jeffrey R Marks, and Joseph R Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462–11467, 2001.

S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

Appendix A. Related Work

Since it became clear that Stochastic Gradient Descent (SGD), with appropriate step size tuning, was an efficient method for solving the training problem (1), there has been a search for an efficient second order counter part. The hope being, and our objective here, is to find a second order stochastic method that is *incremental*; that is, it can work with mini-batches, requires little to *no tuning* since it would depend less on how well scaled or conditioned the data is, and finally, would also apply to *non-convex* problems. To date there is a vast literature on stochastic second order methods, yet none that achieve all of the above.

The subsampled Newton methods such as (Roosta-Khorasani and Mahoney, 2019; Bollapragada et al., 2018; Liu and Roosta, 2021; Erdogdu and Montanari, 2015; Kohler and Lucchi, 2017; Jahani et al., 2017) require large batch sizes in order to guarantee that the subsampled Newton direction is close to the full Newton direction in high probability. As such are not incremental. Other examples of large sampled based methods include the Stochastic quasi-Newton methods (Byrd et al., 2011; Mokhtari and Ribeiro, 2015; Moritz et al., 2016; Gower et al., 2016; Wang et al., 2017; Berahas et al., 2016), stochastic cubic Newton Tripuraneni et al. (2018), SDNA (Qu et al., 2016), Newton sketch (Pilanci and Wainwright, 2017) and Lissa (Agarwal et al., 2017), since these require a large mini-batch or full gradient evaluations.

The only incremental second order methods we aware of are *IQN* (Incremental Quasi-Newton) (Mokhtari et al., 2018), *SNM* (Stochastic Newton Method) (Kovalev et al., 2019; Rodomanov and Kropotov, 2016) and very recently *SAN* (Stochastic Average Newton) Chen et al. (2021). IQN and SNM enjoy a fast local convergence, but their computational and memory costs per iteration, is of $\mathcal{O}(d^2)$ making them prohibitive in large dimensions.

Handling non-convexity in second order methods is particularly challenging because most second order methods rely on convexity in their design. For instance, the classic Newton iteration is the minima of the local quadratic approximation if this approximation is convex. If it is not convex, the Newton step can be meaningless, or worse, a step uphill. Quasi-Newton methods maintain positive definite approximation of the Hessian matrix, and thus are also problematic when applied to non-convex problems Wang et al. (2017) for which the Hessian is typically indefinite. Furthermore the incremental Newton methods IQN, SNM and SAN methods rely on the convexity of f_i in their design. Indeed, without convexity, the iterates of IQN, SNM and SAN are not well defined.

In contrast, our approach of finding roots of the local quadratic approximation (7) in no way relies on convexity, and relies solely on the fact that the local quadratic approximation around w^t is good if we are not far from w^t. But our approach does introduce a new problem: the need to solve a system of quadratic equations. We propose a series of methods to solve this in Sections 2 and C.

Solving quadratic equations has been heavily studied. There are even dedicated methods for solving

$$w^\star = \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2}\|w-\bar{w}\|^2 \;\text{ s.t. }\; Q(w)=0, \;\text{ where }\; Q(w) = \frac{1}{2}w^\top \mathbf{H} w + b^\top w + c \tag{19}$$

for a given $\bar{w}$, where $\mathbf{H}$ is a nonzero symmetric (*not necessarily PSD*) matrix, and the level set $\{w : Q(w) = 0\}$ is nonempty. Note that since $Q(w)$ is a quadratic function, the problem (19) is nonconvex. Yet despite this non-convexity, so long as there exists a feasible point, the projection (19) can be solved in polynomial time by re-writing the projection as a semi-definite program, or by using the S-procedure, which involves computing the eigenvalue decomposition of $\mathbf{H}$ and using a line search as proposed in Park and Boyd (2017), and detailed here in Section D. But this approach is too costly when the dimension d is large.

An alternative iterative method is proposed in Sosa and MP Raupp (2020), but only asymptotic convergence is guaranteed. In Dai (2006), the authors consider a similar problem by projecting a point onto a general ellipsoid, which is again a problem of solving quadratic equations. However, they require the matrix $\mathbf{H}$ to be a positive definite matrix.

The problem (7) and (19) are also an instance of a quadratic constrained quadratic program (QCQP). Although the QCQP in (7) has no closed form solution in general, we show in the next section that there is a closed form solution for Generalized linear models (GLMs), that holds for convex and non-convex GLMs alike. For general non-linear models we propose in Section 2.2 an approximate solution to (7) by iteratively linearizing the quadratic constraint and projecting onto the linearization.

Appendix B. Convergence Theory

Here we provide a convergence theory for SP2 and SP2$^+$ for when $f(w)$ is an average of quadratic functions. Fix $w^* \in \mathbb{R}^d$ and let the loss over the i-th data point be given by

$$f_i(w) = \langle \mathbf{H}_i(w - w^*), w - w^* \rangle \,, \tag{20}$$

where $\mathbf{H}_i \in \mathbb{R}^{d\times d}$ is a symmetric positive semi-definite matrix for $i = 1, \ldots, n$. Consequently $f_i(w^*) = 0 = f(w^*) = \min_{w\in\mathbb{R}^d} f(w)$, thus the interpolation condition holds.

Proposition 5 *Consider the loss functions given in (20). The* `SP2` *method (7) converges linearly*

$$\mathbb{E}\left[\left\|w^{t+1} - w^*\right\|^2\right] \leq \rho\, \mathbb{E}\left[\left\|w^t - w^*\right\|^2\right], \quad \text{where } \rho = \lambda_{\max}\left(\mathbf{I} - \frac{1}{n}\sum_{i=1}^n \mathbf{H}_i\mathbf{H}_i^+\right) < 1. \tag{21}$$

The rate of convergence of `SP2` in (21) can be orders of magnitude better than SGD. Indeed, since (20) is convex, smooth and interpolation holds, we have from Gower et al. (2021b) that SGD converges at a rate of

$$\rho_{SGD} = 1 - \frac{1}{2n}\frac{\lambda_{\min}(\sum_{i=1}^n \mathbf{H}_i)}{\max_{i=1,\ldots,n}\lambda_{\max}(\mathbf{H}_i)}. \tag{22}$$

To compare (22) to ρ rate in Proposition 5, consider the case where all $\mathbf{H}_i$ are invertible. In this case $\mathbf{H}_i\mathbf{H}_i^+ = \mathbf{I}$ and thus $\rho = 0$ and `SP2` converges in one step. Indeed, even if a single $\mathbf{H}_i$ is invertible, after sampling i the `SP2` will converge. In contrast, the SGD method is still at the mercy of the spectra of the $\mathbf{H}_i$ matrices and depend on how well conditioned these matrices are. Even in the extreme case where all $\mathbf{H}_i$ are well conditioned, for example $\mathbf{H}_i = i \times \mathbf{I}$, the rate of convergence of SGD can be very slow, for instance in this case we have $\rho_{SGD} = 1 - \frac{1}{2n^2}$.

Proposition 6 *Consider the loss functions in (20). The* SP2$^+$ *method (18) converges linearly*

$$\mathbb{E}\left[\left\|w^{t+1}-w^*\right\|^2\right] \leq \rho_{SP2+}^2\, \mathbb{E}\left[\left\|w^{t}-w^*\right\|^2\right], \text{ where } \rho_{SP2+} = 1-\frac{1}{2n}\sum_{i=1}^n \frac{\lambda_{\min}(\mathbf{H}_i)}{\lambda_{\max}(\mathbf{H}_i)}. \quad (23)$$

The rate of convergence of SP2$^+$ now depends on the average condition number of the $\mathbf{H}_i$ matrix. Yet still, the rate of convergence in (23) is always better than that of SGD. Indeed, this follows because of the maximum over the index i in (22) and since

$$\frac{1}{2n}\sum_{i=1}^n \frac{\lambda_{\min}(\mathbf{H}_i)}{\lambda_{\max}(\mathbf{H}_i)} \geq \frac{1}{2n}\sum_{i=1}^n \frac{\lambda_{\min}(\mathbf{H}_i)}{\max_{i=1,\ldots,n}\lambda_{\max}(\mathbf{H}_i)}.$$

Note also that the rate of SP2$^+$ appears squared in (23) and the rate ρ_{SGD} of SGD is not squared. But this difference accounts for the fact that each step of SP2$^+$ is at least twice the cost of SGD, since each step of SP2$^+$ is comprised of two gradient steps, see (15) and (17). Thus we can neglect the apparent advantage of the rate ρ_{SP2+} being squared.

Appendix C. Quadratic with Slack

Here we depart from the interpolation Assumption 1 and design a variant of SP2$^+$ that can be applied to models that are *close* to interpolation. Instead of trying to set all the losses to zero, we now will try to find the smallest *slack variable* $s > 0$ for which

$$f_i(w) \leq s, \qquad \text{for } i = 1,\ldots,n.$$

If interpolation holds, then $s = 0$ is a solution. Outside of interpolation, s may be non-zero.

C.1. L2 slack formulation

To make s as small as possible, we can solve the following problem

$$\min_{s\in\mathbb{R}, w\in\mathbb{R}^s} \frac{1}{2}s^2 \text{ subject to } f_i(w) \leq s, \text{ for } i = 1,\ldots,n, \quad (24)$$

which is called the *L2 slack formulation.* This type of slack problem was introduced in Crammer et al. (2006) to derive variants of the passive-aggressive method that could be applied to linear models on non-separable data, in other words, that do not interpolate the data.

To solve (24) we will again project onto a local quadratic approximation of the constraint. Let

$$q_{i,t}(w) \coloneqq f_i(w^t) + \left\langle \nabla f_i(w^t), w-w^t\right\rangle + \tfrac{1}{2}\left\langle \nabla^2 f_i(w^t)(w-w^t), w-w^t\right\rangle. \quad (25)$$

and let $\Delta_t = \left\|w-w^t\right\|^2 + (s-s^t)^2$. Consider the iterative method given by

$$w^{t+1}, s^{t+1} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1-\lambda}{2}\Delta_t + \frac{\lambda}{2}s^2 \qquad \text{s.t. } q_{i,t}(w) \leq s, \quad (26)$$

where $\lambda \in [0,\ 1]$ is a regularization parameter that trades off between having a small s, and using the previous iterates as a regularizor. The resulting projection problem in (26) has a quadratic inequality, and thus in most cases has no closed form solution, despite always being feasible[4].

So instead of solving (26) exactly, we propose an approximate solution by iteratively linearizing and projecting onto the constraints. Our approximate solution has two steps, the first step being

$$w^{t+1/2}, s^{t+1/2} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1-\lambda}{2}\Delta_t + \frac{\lambda}{2}s^2 \quad \text{s.t. } q_{i,t}(w^t) + \left\langle \nabla q_{i,t}(w^t), w - w^t \right\rangle \leq s. \tag{27}$$

The second step is given by projecting $w^{t+1/2}$ onto the linearization around $w^{t+1/2}$ as follows

$$w^{t+1}, s^{t+1} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1-\lambda}{2}\Delta_{t+\frac{1}{2}} + \frac{\lambda}{2}s^2 \ \text{ s.t. } q_{i,t}(w^{t+1/2}) + \left\langle \nabla q_{i,t}(w^{t+1/2}), w - w^{t+1/2} \right\rangle \leq s. \tag{28}$$

The closed form solution to our two step method is given in Lemma 15 of Appendix F.7. We refer to this method as SP2L2$^+$.

C.2. L1 slack formulation

To make s as small as possible, we can also solve the following *L1 slack formulation*

$$\min_{s\geq 0, w\in\mathbb{R}^d} s \quad \text{s.t. } f_i(w) \leq s, \text{ for } i = 1, \ldots, n.$$

Similarly, we can again project onto a local quadratic approximation of the constraint and consider the iterative method given by

$$w^{t+1}, s^{t+1} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1-\lambda}{2}\Delta_t + \frac{\lambda}{2}s \quad \text{s.t. } q_{i,t}(w) \leq s, \tag{29}$$

where $\lambda \in [0,\ 1]$ is a regularization parameter that trades off between having a small s, and using the previous iterates as a regularizor.

To approximately solve (29), we again propose an approximate two step method similar to (27) and (28). The closed form solution to the two step method is given in Lemma 17 of Appendix F.8. We refer to this method as SP2L1$^+$.

C.3. Dropping the Slack Regularization

Note that the objective function in (29) contains a regularization term $(s - s^t)^2$, which forces s to be close to s^t. If we allow s to be far from s^t, we can instead solve the following unregularized problem

$$w^{t+1}, s^{t+1} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1-\lambda}{2}\left\|w - w^t\right\|^2 + \frac{\lambda}{2}s \quad \text{s.t. } q_{i,t}(w) \leq s, \tag{30}$$

4. For instance $w = w^t$ and $s = f_i(w^t)$ is feasible

where $\lambda \in [0,\ 1]$ is again a regularization parameter that trades off between having a small s, and using the previous iterates as a regularizor. We call the resulting method in (30) the `SP2max` method since it is a second order variant of the `SPmax` method Loizou et al. (2020); Gower et al. (2022). The advantage of `SP2max` is that it has a closed form solution for GLMs (8) as shown in the following lemma which is proved in Appendix F.10.

Lemma 7 *(`SP2max`) Consider the GLM model given in (8) and (9). If the loss $f_i = f_i(w^t)$ is non-negative, then the iterates of (30) have a closed form solution given by*

$$w^{t+1} = w^t + c^\star x_i, \qquad s^{t+1} = \max\left\{\widetilde{s},\ 0\right\},$$

and where $\widetilde{s} = f_i - \frac{\widetilde{\lambda} a_i^2 \ell}{1+\widetilde{\lambda} h_i \ell} + \frac{h_i \widetilde{\lambda}^2 a_i^2 \ell^2}{2(1+\widetilde{\lambda} h_i \ell)^2}$, $\ell = \|x_i\|^2$, $\widetilde{\lambda} = \frac{\lambda}{2(1-\lambda)}$, *and*

$$c^\star = \begin{cases} 0, & \text{if } f_i = 0 \\ -\frac{\widetilde{\lambda} a_i}{1+\widetilde{\lambda} h_i \ell}, & \text{if } f_i > 0 \text{ and } \widetilde{s} \geq 0, \\ \frac{-a_i + \sqrt{a_i^2 - 2h_i f_i}}{h_i \ell}, & \text{otherwise.} \end{cases}$$

To approximately solve (30) in general, we again propose an approximate two step method. The closed form solution to the two step method is given in Lemma 18 of Appendix F.9. We refer to this method as $\texttt{SP2max}^+$.

Appendix D. Projecting onto Quadratic

This following projection lemma is based on Section B in Park and Boyd (2017). What we do in addition to Park and Boyd (2017) is to clarify how to compute the resulting projection, and add further details on the proof.

Lemma 8 *Let $w \in \mathbb{R}^d$ and let $\mathbf{P} \in \mathbb{R}^{d\times d}$ is a symmetric matrix. Consider the projection*

$$\begin{aligned} w' \in \operatorname{argmin}_{w\in\mathbb{R}^d} & \tfrac{1}{2}\left\|w - z\right\|^2 \\ \text{s.t. } & r + \langle q, w-z\rangle + \tfrac{1}{2}\langle \mathbf{P}(w-z), w-z\rangle = 0. \end{aligned} \tag{31}$$

Let

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$$

be the eigenvalues of $\mathbf{P}$ and let $Q\Lambda Q^\top = \mathbf{P}$ be the eigenvalue decomposition of $\mathbf{P}$, where $\Lambda = diag(\lambda_i)$ and $QQ^\top = I$. Let $\hat{q} = Q^\top q$. If the quadratic constraint in (31) is feasible, then there exists a solution to (31). Now we give the three candidate solutions.

1. *If $r = 0$, then the solution is given by*

$$w = z.$$

2. *Now assuming $r \neq 0$. Let*

$$\nu = \max_{i\,:\,\lambda_i \neq 0} \left\{-\frac{1}{\lambda_1}, -\frac{1}{\lambda_d}\right\} \tag{32}$$

$$i^* \in \arg\max_{i\,:\,\lambda_i \neq 0} \left\{-\frac{1}{\lambda_1}, -\frac{1}{\lambda_d}\right\} \tag{33}$$

$$N = \{i \ :\ \lambda_i \neq \lambda_{i^*}\}. \tag{34}$$

Let

$$x^* = -(\mathbf{I} + \nu\Lambda)^\dagger \nu \hat{q}. \tag{35}$$

If

$$2\nu r + \nu \langle \hat{q}, x^* \rangle - \|x^*\|^2 + \frac{\nu^2}{4} \sum_{i \in N} \hat{q}_i^2 \geq 0 \tag{36}$$

then the solution is given by

$$w' = z + Q(x^* + n), \tag{37}$$

where $n \in \mathbb{R}^d$ *and*

$$\begin{aligned} n_i &= \frac{\nu}{2}\hat{q}_i + \frac{1}{\sqrt{|N|}} \sqrt{2\nu r + \nu \langle \hat{q}, x^* \rangle - \|x^*\|^2 + \frac{\nu^2}{4} \sum_{i \in N} \hat{q}_i^2}, \quad \textit{for } i \in N \\ n_i &= 0, \quad \textit{for } i \notin N. \end{aligned}$$

3. *Alternatively if* (36) *does not hold, then the solution is given by*

$$w' = z - (\mathbf{I} + \nu\Lambda)^\dagger \nu q \tag{38}$$

where ν *is the solution to the nonlinear equation*

$$\frac{\nu}{2} \sum_i \frac{\hat{q}_i^2 (2 + \nu\lambda_i)}{(1 + \nu\lambda_i)^2} = r. \tag{39}$$

Proof First note that there exists a solution to (31) since the constraint is a closed feasible set.

Let $Q\Lambda Q^\top = \mathbf{P}$ be the SVD of $\mathbf{P}$, where $QQ^\top = I$. By changing variables $x = Q^\top(w - z)$ we have that (31) is equivalent to

$$\begin{aligned} &\mathrm{argmin}_{\hat{x} \in \mathbb{R}^d} \tfrac{1}{2} \|x\|^2 \\ &\text{s.t. } r + \langle \hat{q}, x \rangle + \tfrac{1}{2} \langle \Lambda x, x \rangle = 0, \end{aligned} \tag{40}$$

where $\hat{x} = Q^\top(w - z)$ and $\hat{q} = Q^\top q$. The Lagrangian of (40) is given by

$$\begin{aligned} L(x, \nu) &= \tfrac{1}{2} \|x\|^2 + \nu\big(r + \langle \hat{q}, x \rangle + \tfrac{1}{2} \langle \Lambda x, x \rangle\big) \\ &= \tfrac{1}{2} x^\top (\mathbf{I} + \nu\Lambda) x + \nu\big(r + \langle \hat{q}, x \rangle\big). \end{aligned} \tag{41}$$

Thus the KKT conditions are given by

$$\nabla_x L(x, \nu) = (\mathbf{I} + \nu\Lambda)x + \nu\hat{q} = 0 \tag{42}$$

$$\nabla_\nu L(x, \nu) = r + \langle \hat{q}, x \rangle + \tfrac{1}{2} \langle \Lambda x, x \rangle = 0. \tag{43}$$

Since we are guaranteed that the projection has a solution, we have that as a necessary condition that the solution satisfies

$$\nabla_x^2 L(x, \nu) = (\mathbf{I} + \nu\Lambda) \succeq 0,$$

see Theorem 12.5 in Wright and Nocedal (1999). Consequently either $(\mathbf{I}+\nu\Lambda)\succ 0$ or $(\mathbf{I}+\nu\Lambda)$ has a zero eigenvalue.

Consider the case where $(\mathbf{I}+\nu\Lambda)\succ 0$. From (42) we have that

$$x = -\nu(\mathbf{I}+\nu\Lambda)^{-1}\hat{q}. \tag{44}$$

Now note that if $\nu = 0$ then $x = 0$ and by the constraint we must have $r = 0$. Otherwise, if $r \neq 0$, then $\nu \neq 0$. Assume now $\nu \neq 0$ and substituting the above into (43) and letting $\Lambda = \mathrm{diag}(\lambda_i)$ gives

$$\begin{aligned}
\nabla_\nu L(x,\nu) &= r + \langle \hat{q}, x\rangle + \tfrac{1}{2\nu}\langle \nu\Lambda x, x\rangle \\
&= r + \langle \hat{q}, x\rangle + \tfrac{1}{2\nu}\langle (\mathbf{I}+\nu\Lambda)x, x\rangle - \frac{1}{2\nu}\|x\|^2 \\
&= r + \frac{1}{2}\langle \hat{q}, x\rangle - \frac{1}{2\nu}\|x\|^2 && \text{Using (42)} \\
&= r - \frac{\nu}{2}\left\langle \hat{q}, (\mathbf{I}+\nu\Lambda)^{-1}\hat{q}\right\rangle - \frac{\nu}{2}\left\|(\mathbf{I}+\nu\Lambda)^{-1}\hat{q}\right\|^2 && \text{Using (44)} \\
&= r - \frac{\nu}{2}\sum_i \left(\frac{\hat{q}_i^2}{1+\nu\lambda_i} + \frac{\hat{q}_i^2}{(1+\nu\lambda_i)^2}\right).
\end{aligned}$$

Thus

$$\frac{\nu}{2}\sum_i \frac{\hat{q}_i^2(2+\nu\lambda_i)}{(1+\nu\lambda_i)^2} \;=\; r. \tag{45}$$

Upon finding the solution ν to the above, we have that our final solution is given by $w' = z + Qx$, that is

$$\begin{aligned}
w' &= z - Q(\mathbf{I}+\nu\Lambda)^{\dagger}\nu\hat{q} \\
&= z - (\mathbf{I}+\nu\Lambda)^{\dagger}\nu Q\hat{q} \\
&= z - \nu(\mathbf{I}+\nu\Lambda)^{\dagger} q
\end{aligned} \tag{46}$$

Alternatively, suppose that $(\mathbf{I}+\nu\Lambda)\succeq 0$ is non-singular. The positive definiteness implies that

$$\nu \;\geq\; -\frac{1}{\lambda_i}, \quad \text{for } i = 1,\ldots,d. \tag{47}$$

For $(\mathbf{I}+\nu\Lambda)$ to be non-singular, at least one of the above inequalities will hold to equality. To ease notation, let us arrange the eigenvalues in increasing order so that

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d.$$

For one of the (47) inequalities to hold to equality we need that

$$\nu \;=\; \max_{i\,:\,\lambda_i\neq 0} -\frac{1}{\lambda_i} = \max_{i\,:\,\lambda_i\neq 0}\left\{-\frac{1}{\lambda_1}, -\frac{1}{\lambda_d}\right\}.$$

Since $(\mathbf{I}+\nu\Lambda)$ is now singular with this ν, we have that the solution to (42) is given by

$$x \;=\; -(\mathbf{I}+\nu\Lambda)^{\dagger}\nu\hat{q} + n \;:=\; x^* + n, \quad \text{where } \langle x^*, n\rangle = 0, \tag{48}$$

where † denotes the pseudo-inverse and where n is in the kernel of $(\mathbf{I}+\nu\Lambda)$, in other words $(\mathbf{I}+\nu\Lambda)n = 0$. It remains to determine n, which we can do with (43). Indeed, substituting (48) into (43) gives

$$\begin{aligned}\nabla_\nu L(x,\nu) &= r + \frac{1}{2}\langle \hat{q}, x\rangle - \frac{1}{2\nu}\|x\|^2 && \text{Using (42)}\\ &= r + \frac{1}{2}\langle \hat{q}, x^*+n\rangle - \frac{1}{2\nu}\|n\|^2 - \frac{1}{2\nu}\|x^*\|^2. && \text{Using (48).}\end{aligned}$$

Setting to zero and completing the squares in n we have that

$$\frac{1}{2\nu}\left\|n - \frac{\nu}{2}\hat{q}\right\|^2 = r + \frac{1}{2}\langle \hat{q}, x^*\rangle - \frac{1}{2\nu}\|x^*\|^2 + \frac{\nu}{8}\|\hat{q}\|^2. \tag{49}$$

To characterize the solutions in n of the above, first note that n will only have a few non-zero elements. To see this, let $i^* \in \arg\max_{i\,:\,\lambda_i\neq 0}\left\{-\frac{1}{\lambda_1}, -\frac{1}{\lambda_d}\right\}$, and note that $(\mathbf{I}+\nu\Lambda)$ has as many zeros on the diagonal as the multiplicity of the eigenvalue λ_{i^*}. That is, it has zeros elements on the indices in

$$I \;=\; \{i \;:\; \lambda_i = \lambda_{i^*}\}.$$

Thus the non-zero elements of n are in the set

$$N \;=\; \{i \;:\; \lambda_i \neq \lambda_{i^*}\}.$$

Because of this observation we further re-write (49) as

$$\begin{aligned}\sum_{i\in N}\left(n_i - \frac{\nu}{2}\hat{q}_i\right)^2 &= 2\nu r + \nu\langle \hat{q}, x^*\rangle - \|x^*\|^2 + \frac{\nu^2}{4}\|\hat{q}\|^2 - \sum_{i\in I}\frac{\nu^2}{4}\hat{q}_i^2\\ &= 2\nu r + \nu\langle \hat{q}, x^*\rangle - \|x^*\|^2 + \frac{\nu^2}{4}\sum_{i\in N}\hat{q}_i^2. \end{aligned}\tag{50}$$

Consequently, if the above is positive, then there exists solutions to the above of which

$$n_i = \frac{\nu}{2}\hat{q}_i + \frac{1}{\sqrt{|N|}}\sqrt{2\nu r + \nu\langle \hat{q}, x^*\rangle - \|x^*\|^2 + \frac{\nu^2}{4}\sum_{i\in N}\hat{q}_i^2}, \quad \text{for } i \in N. \tag{51}$$

is one. Consequently, the final solution is given by $w = z + Q(x^*+n)$ where x^* is given in by (48). ∎

Corollary 9 *If $r > 0$ and $\mathbf{P}$ has at least one negative eigenvalue, there always exists a solution to the projection* (31).

Proof We only need to prove that there exists a solution to the quadratic equation in (31), after which Lemma 8 guarantees the existance of a solution. ∎

Appendix E. Matrix Completion

The projection (117) can be solved as we shown in the following theorem.

Theorem 10 *The solution to* (117) *is given by one of the following cases.*

1. *If* $(u_j^k)^\top v_j^k = a_{i,j}$ *then* $u = u_i^k$ *and* $v = v_j^k$.
2. *Alternatively if* $u_j^k = v_j^k$ *and*
$$(u_j^k)^\top v_j^k \geq 4a_{i,j}$$
then
$$v = \frac{1}{2}v_j^k + \frac{1}{2}\frac{v_j^k}{\left\|v_j^k\right\|}\sqrt{\left\|v_j^k\right\|^2 - 4a_{i,j}} \tag{52}$$
$$u = -\frac{1}{2}u_j^k + \frac{1}{2}\frac{u_j^k}{\left\|u_j^k\right\|}\sqrt{\left\|u_j^k\right\|^2 - 4a_{i,j}} \tag{53}$$
3. *Finally, if none of the above holds then*
$$u = \frac{u_i^k - \gamma v_j^k}{1-\gamma^2} \tag{54}$$
$$v = \frac{v_j^k - \gamma u_i^k}{1-\gamma^2}, \tag{55}$$
where $\gamma \in (-1, 1)$ *and is the solution to the depressed quartic equation*
$$(1+\gamma^2)\left\langle u_i^k, v_j^k\right\rangle - \gamma(\left\|u_i^k\right\|^2 + \left\|v_j^k\right\|^2) = (1-\gamma^2)^2 a_{i,j}. \tag{56}$$

Proof The Lagrangian of (117) is given by

$$L(u, v, \gamma) \ = \ \tfrac{1}{2}\left\|u - u_i^k\right\|^2 + \tfrac{1}{2}\left\|v - v_j^k\right\|^2 + \gamma(u^\top v - a_{i,j}),$$

where $\gamma \in \mathbb{R}$ is the unknown Lagrangian Multiplier. Thus the KKT equations are

$$u - u_i^k + \gamma v = 0 \tag{57}$$
$$v - v_j^k + \gamma u = 0 \tag{58}$$
$$u^\top v = a_{i,j} \tag{59}$$

Subtracting γ times the the second equation from the first equation, and analogously, subtracting the first equation from the second gives

$$(1-\gamma^2)u - u_i^k + \gamma v_j^k = 0 \tag{60}$$
$$(1-\gamma^2)v - v_j^k + \gamma u_i^k = 0. \tag{61}$$

If $\gamma = 1$, then necessarily $u_i^k = v_j^k$ and furthermore from the first equation in (59) we have that

$$u = u_i^k - v = v_j^k - v. \tag{62}$$

Substituting u out in the original projection problem (117) we have that

$$\min_v \|v\|^2 - v^\top v_j^k \ \text{ subject to } \ v^\top v_j^k - \|v\|^2 = a_{i,j}. \tag{63}$$

Consequently, for every v that satisfies the constraint we have that the objective value is invariant and equal to $-a_{i,j}$. Consequently there are infinite solutions. To find one such solution, we complete the squares of the constraint and find

$$\left\|v - \tfrac{1}{2}v_j^k\right\|^2 \quad = \quad \frac{1}{4}\left\|v_j^k\right\|^2 - a_{i,j}. \tag{64}$$

The above only has solutions if $\frac{1}{4}\left\|v_j^k\right\|^2 - a_{i,j} \geq 0$. One solution to the above is given by (52).

Alternatively, if $\gamma \neq 1$ then by isolating u and v in (60) and (61), respectively, gives

$$u = \frac{u_i^k - \gamma v_j^k}{1-\gamma^2} \tag{65}$$

$$v = \frac{v_j^k - \gamma u_i^k}{1-\gamma^2} \tag{66}$$

To figure out γ, we use the third constraint in (59) and the above two equations, which gives

$$\begin{aligned} u^\top v &= \frac{(u_i^k - \gamma v_j^k)^\top}{1-\gamma^2}\frac{v_j^k - \gamma u_i^k}{1-\gamma^2} \\ &= \frac{(1+\gamma^2)\left\langle u_i^k, v_j^k\right\rangle - \gamma(\|u_i^k\|^2 + \left\|v_j^k\right\|^2)}{(1-\gamma^2)^2} \quad = \quad a_{i,j}. \end{aligned}$$

Let

$$\phi(\gamma) = (1+\gamma^2)\left\langle u_i^k, v_j^k\right\rangle - \gamma(\left\|u_i^k\right\|^2 + \left\|v_j^k\right\|^2) - (1-\gamma^2)^2 a_{i,j}.$$

Can we now find an interval which will contain the solution in γ? Note that

$$\begin{aligned} \phi(-1) &= 2\left\langle u_i^k, v_j^k\right\rangle + \left\|u_i^k\right\|^2 + \left\|v_j^k\right\|^2 = \left\|u_i^k + v_j^k\right\|^2 \geq 0 \\ \phi(1) &= 2\left\langle u_i^k, v_j^k\right\rangle - \left\|u_i^k\right\|^2 - \left\|v_j^k\right\|^2 = -\left\|u_i^k - v_j^k\right\|^2 \leq 0. \end{aligned}$$

Thus it suffices to search for $\gamma \in (-1,\ 1)$, which can be done efficiently with bisection.

∎

Appendix F. Proofs of Important Lemmas

F.1. Proof of Lemma 2

Let us first describe the set of solutions for given constraint. We need to have

$$f_i + a_i x_i^T \Delta + \frac{1}{2} h_i \Delta^T x_i x_i^T \Delta = 0, \tag{67}$$

where $\Delta = w - w^t$ is unknown. If we denote by $\tau_i = x_i^T \Delta$ then (67) will reduce to

$$f_i + a_i \tau_i + \frac{1}{2} h_i \tau_i^2 = 0. \tag{68}$$

This quadratic equation (68) has solution if

$$a_i^2 - 2h_i f_i \geq 0. \tag{69}$$

If the condition above holds then we have that the solution for τ is in this set

$$T^* := \left\{ \frac{-a_i + \sqrt{a_i^2 - 2h_i f_i}}{h_i}, \frac{-a_i - \sqrt{a_i^2 - 2h_i f_i}}{h_i} \right\}. \tag{70}$$

Recall that the problem (7) now reduces into

$$\min_{\Delta} \left\{ \|\Delta\|^2, \text{ such that } x_i^T \Delta \in T^* \right\}. \tag{71}$$

Note that because we want to minimize $\|\Delta\|^2$, we want to choose the constraint with smallest possible absolute value, hence the problem (71) is equivalent to

$$\min_{\Delta} \left\{ \|\Delta\|^2, \text{ such that } x_i^T \Delta = \tau_i^* \right\}, \tag{72}$$

where

$$\tau_i^* = \begin{cases} \frac{-a_i + \sqrt{a_i^2 - 2h_i f_i}}{h_i}, & \text{if } a_i > 0 \\ \frac{-a_i - \sqrt{a_i^2 - 2h_i f_i}}{h_i}, & \text{otherwise.} \end{cases}$$

In other words,

$$\tau_i^* = -\frac{a_i}{h_i} + \text{sign}(a_i) \frac{\sqrt{a_i^2 - 2h_i f_i}}{h_i} = \frac{a_i}{h_i} \left(\frac{\sqrt{a_i^2 - 2h_i f_i}}{|a_i|} - 1 \right)$$

The final solution is hence

$$\Delta^* = \frac{\tau_i^*}{\|x_i\|^2} x_i$$

and therefore

$$w^* = w^t + \frac{\tau_i^*}{\|x_i\|^2} x_i$$

In case when (69) is not satisfied, and because we assumed that the loss function is non-negative, we necessary have $h_i > 0$. Then natural choice of τ_i is the one that would minimize the

$$f_i + a_i\tau_i + \frac{1}{2}h_i\tau_i^2.$$

From first order optimality conditions we obtain that

$$\tau_i^* = -\frac{a_i}{h_i}$$

which leads to (12).

F.2. Proof of Lemma 3

Proof If $\phi(t) = 0$ then the condition holds trivially. For t such that $\phi(t) \neq 0$, $\sqrt{\phi(t)}$ is differentiable, and we have

$$\frac{d^2}{dt^2}\sqrt{\phi(t)} = -\tfrac{1}{4}\phi(t)^{-3/2}\phi'(t)^2 + \tfrac{1}{2}\phi(t)^{-1/2}\phi''(t) = \tfrac{1}{4}\phi(t)^{-3/2}(-\phi'(t)^2 + 2\phi(t)\phi''(t)),$$

which is negative precisely when $\phi'(t)^2 \geq 2\phi(t)\phi''(t)$. ∎

F.3. Proof of Lemma 4

Note that

$$\begin{aligned} q(w^{t+1/2}) &\overset{(15)+(13)}{=} f_i(w^t) - \left\langle \nabla f_i(w^t), \frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t)\right\rangle \\ &\quad + \frac{1}{2}\left\langle \nabla^2 f_i(w^t)\frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t), \frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t)\right\rangle \\ &= \frac{1}{2}\frac{f_i(w^t)^2}{\|\nabla f_i(w^t)\|^4}\left\langle \nabla^2 f_i(w^t)\nabla f_i(w^t), \nabla f_i(w^t)\right\rangle. \end{aligned}$$

Furthermore

$$\begin{aligned} \nabla q(w^{t+1/2}) &\overset{(13)}{=} \nabla f_i(w^t) + \nabla^2 f_i(w^t)(w^{t+1/2} - w^t) \\ &\overset{(15)}{=} \left(\mathbf{I} - \nabla^2 f_i(w^t)\frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\right)\nabla f_i(w^t). \end{aligned}$$

Thus the second step (17) is given by

$$w^{t+1} = w^{t+1/2} - \frac{1}{2}\frac{f_i(w^t)^2}{\|\nabla f_i(w^t)\|^4}\frac{\left\langle \nabla^2 f_i(w^t)\nabla f_i(w^t), \nabla f_i(w^t)\right\rangle}{\left\|\left(\mathbf{I} - \nabla^2 f_i(w^t)\frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\right)\nabla f_i(w^t)\right\|^2}\cdot\left(\mathbf{I} - \nabla^2 f_i(w^t)\frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\right)\nabla f_i(w^t). \tag{73}$$

Putting the first (15) and second (73) updates together gives (18).

This gives a second order correction of the Polyak step that only requires computing a single Hessian-vector product that can be done efficiently using an additional backwards pass of the function. We call this method `SP2`.

F.4. Convergence of multi-step SP2$^+$

If we apply multiple steps of the SP2$^+$, as opposed to two steps, the method converges to the solution of (13). This follows because each step of SP2$^+$ is a step of NR Newton Raphson's method applied to solving the nonlinear equation

$$q(w) \coloneqq f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle + \tfrac{1}{2} \left\langle \nabla^2 f_i(w^t)(w - w^t), w - w^t \right\rangle .$$

Indeed, starting from $w^0 = w^t$, the iterates of the NR (Newton Raphson) method are given by

$$w^{i+1} = w^i - \left(\nabla q(w^i)^\top \right)^\dagger q(w^i) = w^i - \frac{q(w^i)}{\|\nabla q(w^i)\|^2} \nabla q(w^i), \tag{74}$$

where $\mathbf{M}^\dagger$ denotes the pseudo-inverse of the matrix $\mathbf{M}$.

The NR iterates in (74) can also be written in a variational form given by

$$\begin{aligned} w^{i+1} = & \operatorname{argmin}_{w \in \mathbb{R}^d} \left\| w - w^i \right\|^2 \\ & \text{s.t. } \; q(w^i) + \nabla q(w^i)(w - w^i) = 0. \end{aligned} \tag{75}$$

Comparing the above to the first (14) and second step (16) are indeed two steps of the NR method. Further, we can see that (75) is indeed the multi-step version of SP2$^+$.

This method (74) is also known as gradient descent with a Polyak Step step size, or SP for short. It is this connection we will use to prove the convergence of (74) to a root of $q(w)$.

We assume that $q(w)$ has at least one root. Let $w_q^* \in \mathbb{R}^d$ be a least norm root of $q(w)$, that is

$$w_q^* = \operatorname{argmin} \|w\|^2 \quad \text{subject to } \; q(w) = 0. \tag{76}$$

It follows from Theorem 3.2 of Sosa and MP Raupp (2020) that the above optimization (76) has solution if and only if the following matrix

$$\begin{aligned} B = & (\nabla f_i(w^t) - \nabla^2 f_i(w^t) w^t)(\nabla f_i(w^t) - \nabla^2 f_i(w^t) w^t)^\top + \\ & 2 \left(-f_i(w^t) + \nabla f_i(w^t)^\top w^t - \frac{1}{2} {w^t}^\top \nabla^2 f_i(w^t) w^t \right) \nabla^2 f_i(w^t) \end{aligned} \tag{77}$$

has at least a non-negative eigenvalue.

Theorem 11 *Assume that the matrix B defined in (77) has at least a non-negative eigenvalue. If $q(w)$ is star-convex with respect to w_q^*, that is if*

$$(w^i - w_q^*)^\top \nabla^2 f_i(w^t)(w^i - w_q^*) \geq 0, \quad \textit{for all } i, \tag{78}$$

then it follows that

$$\min_{i=0,\ldots,T-1} q(x^i) \; \leq \; \frac{\sigma_{\max}(\nabla^2 f_i(w^t))}{2T} \left\| w^0 - w_q^* \right\|^2 . \tag{79}$$

Proof The proof follows by applying the convergence Theorem 4.4 in Gower et al. (2021b) or equivalently Corollary D.3 in Gower et al. (2021a). This result first appeared in Theorem 4.4 in Gower et al. (2021b), but we apply Corollary D.3 in Gower et al. (2021a) since it is a bit simpler.

To apply this Corollary D.3 in Gower et al. (2021a), we need to verify that q is an L–smooth function and star-convex. To verify if it is smooth, we need to find $L > 0$ such that

$$q(w) \;\leq\; q(y) + \langle \nabla q(y), w - y\rangle + \frac{L}{2}\left\|w - y\right\|^2, \tag{80}$$

which holds with $L = \sigma_{\max}(\nabla^2 q(y)) = \sigma_{\max}(\nabla^2 f_i(w^t)$ since q is a quadratic function. Furthermore, for q to be star-convex along the iterates w^i, we need to verify if

$$q(w_q^*) \;\geq\; q(w^i) + \left\langle \nabla q(w^i), w^* - w^i \right\rangle. \tag{81}$$

Since q is a quadratic, we have that

$$q(w_q^*) = q(w^i) + \left\langle \nabla q(w^i), w^* - w^i \right\rangle + \left\langle \nabla^2 q(w^i)(w^* - w^i), w^* - w^i \right\rangle.$$

Using this in (82) gives that

$$0 \;\geq\; \left\langle \nabla^2 q(w^i)(w^* - w^i), w^* - w^i \right\rangle = \left\langle \nabla^2 f_i(w^t)(w^* - w^i), w^* - w^i \right\rangle, \tag{82}$$

which is equivalent to our assumption (78). We can now apply the result in Corollary D.3 in Gower et al. (2021a) which states that

$$\min_{i=0,\ldots,T-1}(q(x^i) - q(w_q^*)) \;\leq\; \frac{L}{2T}\left\|w^0 - w_q^*\right\|^2.$$

Finally using $q(w_q^*) = 0$ and that $L = \sigma_{\max}(\nabla^2 f_i(w^t))$ gives the result. ∎

To simplify notation, we will omit the dependency on w^t and denote $c = f_i(w^t)$, $g = \nabla f_i(w^t)$ and $\mathbf{H} = \nabla^2 f_i(w^t)$, thus

$$\begin{aligned} q(w) &= c + \left\langle g, w - w^t \right\rangle + \frac{1}{2}\left\langle \mathbf{H}(w - w^t), w - w^t \right\rangle \\ \nabla q(w) &= g + \mathbf{H}(w - w^t) \\ \nabla^2 q(w) &= \mathbf{H} \end{aligned} \tag{83}$$

Lemma 12 *If $g \in \mathrm{Range}(\mathbf{H})$ and $w^0 \in \mathrm{Range}(\mathbf{H})$ then $w^i, \nabla q(w^i) \in \mathrm{Range}(\mathbf{H})$ for all i and $w_q^* \in \mathrm{Range}(\mathbf{H})$.*

Proof First, note that since $g \in \mathrm{Range}(\mathbf{H})$ and since $\nabla q(w) = q + \mathbf{H}(w - w^t)$ (see (83)) we have that $\nabla q(w) \in \mathrm{Range}(\mathbf{H})$ for all w. Consequently by induction if $w^i \in \mathrm{Range}(\mathbf{H})$ then by (74) we have that $w^{i+1} \in \mathrm{Range}(\mathbf{H})$ since it is a combination of $\nabla q(w^i)$ and w^i.

Finally, let $w_q^* = w^t + w_{\mathbf{H}} + w_{\mathbf{H}}^{\perp}$ where $w_{\mathbf{H}} \in \mathrm{Range}(\mathbf{H})$ and $w_{\mathbf{H}}^{\perp} \in \mathrm{Range}(\mathbf{H})^{\perp}$. It follows that

$$q(w_q^*) = q(w^t + w_{\mathbf{H}}).$$

Furthermore, by orthogonality and Pythagoras' Theorem

$$\|w_q^*\| = \|w^t + w_{\mathbf{H}}\| + \left\|w_{\mathbf{H}}^{\perp}\right\|$$

Consequently, since w_q^* is the least norm solution, we must have that $w_{\mathbf{H}}^{\perp} = 0$ and thus $w_q^* \in \mathrm{Range}(\mathbf{H})$. ■

F.5. Proof of Proposition 5

First we repeat the proposition for ease of reference.

Proposition 13 *Consider the loss functions given in (20). The SP2 method converges (7) converges according to*

$$\mathbb{E}\left[\left\|w^{t+1} - w^*\right\|^2\right] \leq \rho\, \mathbb{E}\left[\left\|w^t - w^*\right\|^2\right], \tag{84}$$

where

$$\rho = \lambda_{\max}\left(\mathbf{I} - \frac{1}{n}\sum_{i=1}^n \mathbf{H}_i \mathbf{H}_i^+\right) \;<\; 1. \tag{85}$$

Proof First consider the first iterate of SP2 which applied to (20) are given by

$$\begin{aligned} w^{t+1} = \min_{w \in \mathbb{R}^d} &\left\|w - w^t\right\|^2 \\ \text{s.t.} &\;\; \|w - w^*\|^2_{\mathbf{H}_i} = 0. \end{aligned}$$

Thus every solution to the constraint set must satisfy

$$w \in w^* + \mathbf{N}_i \alpha, \tag{86}$$

where $\mathbf{N}_i \in \mathbb{R}^{d \times d}$ is a basis for the null space of $\mathbf{H}_i$. where $\alpha \in \mathbb{R}^d$. Substituting into the objective we have the resulting linear least squares problem given by

$$\min_{\alpha \in \mathbb{R}^d} \left\|w^* + \mathbf{N}_i \alpha - w^t\right\|^2$$

The minimal norm solution in α is thus

$$\alpha = \mathbf{N}_i^+ (w^t - w^*)$$

which when substituted into (86) gives

$$w^{t+1} \;\; = \;\; w^* + \mathbf{N}_i \mathbf{N}_i^+ (w^t - w^*). \tag{87}$$

Note that $\mathbf{P}_i \coloneqq \mathbf{N}_i \mathbf{N}_i^+$ is the orthogonal projector onto $\mathrm{Null}(\mathbf{H}_i)$. Subtracting w^* from both sides of (87) and applying the squared norm we have that

$$\begin{aligned} \left\|w^{t+1} - w^*\right\|^2 &= \left\|\mathbf{P}_i (w^t - w^*)\right\|^2 \\ &= \left\langle \mathbf{P}_i (w^t - w^*), (w^t - w^*)\right\rangle \end{aligned} \tag{88}$$

where we used that $\mathbf{P}_i\mathbf{P}_i = \mathbf{P}_i$ because it is a projection matrix. Now taking expectation conditioned on w^t we have

$$\begin{aligned}\mathbb{E}\left[\left\|w^{t+1}-w^*\right\|^2 \mid w^t\right] &= \left\langle \mathbb{E}\left[\mathbf{P}_i\right](w^t-w^*), (w^t-w^*)\right\rangle \\ &\leq \lambda_{\max}(\mathbb{E}\left[\mathbf{P}_i\right])\left\|w^t-w^*\right\|^2.\end{aligned}$$

Since the null space is orthogonal to the range of adjoint, we have that

$$\mathbf{P}_i = \mathbf{I} - \mathbf{H}_i\mathbf{H}_i^+.$$

Thus taking expectation again gives the result (84).

Finally, the rate of convergence ρ in (85) is always smaller than one because, due Jensen's inequality and that $\lambda_{\max}$ is convex over positive definite matrices we have that

$$0 < \lambda_{\max}(\mathbb{E}\left[\mathbf{H}_i\mathbf{H}_i^*\right]) \leq \mathbb{E}\left[\lambda_{\max}(\mathbf{H}_i\mathbf{H}_i^*)\right] = 1, \tag{89}$$

where the greater than zero follows since there must exist $\mathbf{H}_i \neq 0$, otherwise the result still holds and the method converges in one step (with $\rho = 0$). Now multiplying (89) by -1 then adding 1 gives

$$1 > \lambda_{\max}(\mathbf{I} - \mathbb{E}\left[\mathbf{H}_i\mathbf{H}_i^*\right]) \geq 0. \tag{90}$$

■

F.6. Proof of Proposition 6

For convenience we repeat the statement of the proposition here.

Proposition 14 *Consider the loss functions given in (20). The* `SP2`$^+$ *method converges (18) converges according to*

$$\mathbb{E}\left[\left\|w^{t+1}-w^*\right\|^2\right] \leq \rho_{SP2+}^2\,\mathbb{E}\left[\left\|w^t-w^*\right\|^2\right], \tag{91}$$

where

$$\rho_{SP2+} = 1 - \frac{1}{2n}\sum_{i=1}^n \frac{\lambda_{\min}(\mathbf{H}_i)}{\lambda_{\max}(\mathbf{H}_i)} \tag{92}$$

Proof The proof follows simply by observing that for quadratic function the `SP2`$^+$ is equivalent to applying two steps of the `SP` method (4). Indeed in Section 2.2 the `SP2`$^+$ applies two steps of the `SP` method to the local quadratic approximation of the function we wish to minimize. But in this case, since our function is quadratic, it is itself equal to it's local quadratic.

Consequently we can apply the convergence theory of `SP` for smooth, strongly convex functions that satisfy the interpolation condition, such as Corollary 5.7.I in Gower et al. (2021a), which states that `SP` converges at a rate of (92) ■

F.7. Proof of Lemma 15

The following lemma gives the two step update for `SP2L2`$^+$.

Lemma 15 *(SP2L2$^+$) The w^{t+1} and s^{t+1} update of (27)–(28) is given by*

$$
\begin{aligned}
w^{t+1} &= w^t - (\Gamma_1 + \Gamma_2)\nabla f_i(w^t) + \Gamma_2\Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t),\\
s^{t+1} &= (1-\lambda)\big((1-\lambda)(s^t + \Gamma_1) + \Gamma_2\big),
\end{aligned}
$$

where $$\Gamma_1 := \frac{(f_i(w^t) - (1-\lambda)s^t)_+}{1-\lambda + \|\nabla f_i(w^t)\|^2},$$

$$\Gamma_2 := \left(\frac{f_i(w^t) - \Gamma_1\|\nabla f_i(w^t)\|^2 - (1-\lambda)^2(s^t+\Gamma_1)}{1-\lambda+\|\nabla f_i(w^t) - \Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t)\|^2} + \frac{\frac{1}{2}\Gamma_1^2\langle\nabla^2 f_i(w^t)\nabla f_i(w^t), \nabla f_i(w^t)\rangle}{1-\lambda+\|\nabla f_i(w^t) - \Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t)\|^2}\right)_+,$$

where we denote $(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise}\end{cases}$.

We will use the following lemma to prove Lemma 15, which has been proven in Lemma C.2 of Gower et al. (2022).

Lemma 16 (L2 Unidimensional Inequality Constraint) *Let $\delta > 0, c \in \mathbb{R}$ and $w, w^0, a \in \mathbb{R}^d$. The closed form solution to*

$$
\begin{aligned}
w', s' = &\operatorname{argmin}_{w\in\mathbb{R}^d, s\in\mathbb{R}^b} \left\|w - w^0\right\|^2 + \delta\left\|s - s^0\right\|^2 \\
&\text{s.t. } a^\top(w - w^0) + c \leq s \ ,
\end{aligned}
\tag{93}
$$

is given by

$$w' = w^0 - \delta\frac{(c - s^0)_+}{1 + \delta\|a\|^2}a, \tag{94}$$

$$s' = s^0 + \frac{(c - s^0)_+}{1 + \delta\|a\|^2}, \tag{95}$$

where we denote $(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise}\end{cases}$.

We are now in the position to prove Lemma 15. Note that

$$
\begin{aligned}
\frac{1-\lambda}{2}(s - s^t)^2 + \frac{\lambda}{2}s^2 &= \frac{1-\lambda}{2}s^2 - (1-\lambda)ss^t + \frac{\lambda}{2}s^2 + \frac{1-\lambda}{2}(s^t)^2 \\
&= \frac{1}{2}s^2 - (1-\lambda)ss^t + \frac{1-\lambda}{2}(s^t)^2 \\
&= \frac{1}{2}(s - (1-\lambda)s^t)^2 + \frac{\lambda - \lambda^2}{2}(s^t)^2.
\end{aligned}
\tag{96}
$$

Consequently (27) is equivalent to

$$
\begin{aligned}
w^{t+1/2}, s^{t+1/2} \quad = \quad &\underset{s\geq 0,\, w\in\mathbb{R}^d}{\operatorname{argmin}} \left\|w - w^t\right\|^2 + \frac{1}{1-\lambda}(s - (1-\lambda)s^t)^2 \\
&\text{s.t. } q_{i,t}(w^t) + \left\langle\nabla q_{i,t}(w^t), w - w^t\right\rangle \leq s.
\end{aligned}
\tag{97}
$$

It follows from Lemma 16 that the closed form solution is

$$w^{t+1/2} = w^t - \frac{1}{1-\lambda}\frac{(q_{i,t}(w^t) - (1-\lambda)s^t)_+}{1 + \frac{1}{1-\lambda}\|\nabla q_{i,t}(w^t)\|^2}\nabla q_{i,t}(w^t), \tag{98}$$

$$s^{t+1/2} = (1-\lambda)s^t + \frac{(q_{i,t}(w^t) - (1-\lambda)s^t)_+}{1 + \frac{1}{1-\lambda}\|\nabla q_{i,t}(w^t)\|^2}, \tag{99}$$

where we denote

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Note that $q_{i,t}(w^t) = f_i(w^t)$ and $\nabla q_{i,t}(w^t) = \nabla f_i(w^t)$. To simplify the notation, we also denote

$$\Gamma_1 = \frac{1}{1-\lambda}\frac{(f_i(w^t) - (1-\lambda)s^t)_+}{1 + \frac{1}{1-\lambda}\|\nabla f_i(w^t)\|^2}.$$

With this notation we have that

$$\begin{aligned} w^{t+1/2} &= w^t - \frac{1}{1-\lambda}\frac{(f_i(w^t) - (1-\lambda)s^t)_+}{1 + \frac{1}{1-\lambda}\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t) && (100)\\ &= w^t - \Gamma_1 \nabla f_i(w^t), && (101)\\ s^{t+1/2} &= (1-\lambda)s^t + \frac{(f_i(w^t) - (1-\lambda)s^t)_+}{1 + \frac{1}{1-\lambda}\|\nabla f_i(w^t)\|^2} && (102)\\ &= (1-\lambda)(s^t + \Gamma_1). && (103) \end{aligned}$$

In a completely analogous way, the closed form solution to (28) is

$$\begin{aligned} w^{t+1} &= w^{t+1/2} - \frac{1}{1-\lambda}\frac{(q_{i,t}(w^{t+1/2}) - (1-\lambda)s^{t+1/2})_+}{1 + \frac{1}{1-\lambda}\left\|\nabla q_{i,t}(w^{t+1/2})\right\|^2} \cdot \nabla q_{i,t}(w^{t+1/2}),\\ s^{t+1} &= (1-\lambda)s^{t+1/2} + \frac{(q_{i,t}(w^{t+1/2}) - (1-\lambda)s^{t+1/2})_+}{1 + \frac{1}{1-\lambda}\left\|\nabla q_{i,t}(w^{t+1/2})\right\|^2}. \end{aligned} \tag{104}$$

Note that

$$\begin{aligned} q_{i,t}(w^{t+1/2}) &= f_i(w^t) - \left\langle \nabla f_i(w^t), \Gamma_1 \nabla f_i(w^t)\right\rangle + \frac{1}{2}\left\langle \nabla^2 f_i(w^t)\Gamma_1\nabla f_i(w^t), \Gamma_1 \nabla f_i(w^t)\right\rangle\\ &= f_i(w^t) - \Gamma_1\left\|\nabla f_i(w^t)\right\|^2 + \frac{1}{2}\Gamma_1^2\left\langle \nabla^2 f_i(w^t)\nabla f_i(w^t), \nabla f_i(w^t)\right\rangle \end{aligned}$$

and

$$\begin{aligned} \nabla q_{i,t}(w^{t+1/2}) &= \nabla f_i(w^t) + \nabla^2 f_i(w^t)(w^{t+1/2} - w^t)\\ &= \nabla f_i(w^t) - \Gamma_1 \nabla^2 f_i(w^t)\nabla f_i(w^t). \end{aligned}$$

Denoting Γ_2 as in the statement of the lemma we conclude that

$$
\begin{aligned}
w^{t+1} &= w^{t+1/2} - \frac{1}{1-\lambda}\frac{(q_{i,t}(w^{t+1/2}) - (1-\lambda)s^{t+1/2})_+}{1+\frac{1}{1-\lambda}\left\|\nabla q_{i,t}(w^{t+1/2})\right\|^2}\cdot \nabla q_{i,t}(w^{t+1/2}) \\
&= w^t - \Gamma_1 \nabla f_i(w^t) - \Gamma_2\left(\nabla f_i(w^t) - \Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t)\right), \qquad (105) \\
s^{t+1} &= (1-\lambda)s^{t+1/2} + \frac{(q_{i,t}(w^{t+1/2}) - (1-\lambda)s^{t+1/2})_+}{1+\frac{1}{1-\lambda}\left\|\nabla q_{i,t}(w^{t+1/2})\right\|^2} \\
&= (1-\lambda)\left((1-\lambda)(s^t+\Gamma_1)+\Gamma_2\right) \qquad (106)
\end{aligned}
$$

F.8. Proof of Lemma 17

The following Lemma gives a closed form for the two-step update for `SP2L1`$^+$.

Lemma 17 *(`SP2L1`$^+$) The w^{t+1} and s^{t+1} update is given by*

$$
\begin{aligned}
w^{t+1} &= w^t - (\Gamma_4+\Gamma_6)\nabla f_i(w^t) + \Gamma_6\Gamma_4\nabla^2 f_i(w^t)\nabla f_i(w^t), \\
s^{t+1} &= \left(\left(s^t - \frac{\lambda}{2(1-\lambda)} + \Gamma_3\right)_+ - \frac{\lambda}{2(1-\lambda)} + \Gamma_5\right)_+,
\end{aligned}
$$

where

$$
\begin{aligned}
\Gamma_3 &= \frac{\left(f_i(w^t) - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1+\|\nabla f_i(w^t)\|^2}, \quad \Gamma_4 = \min\left\{\Gamma_3, \frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\right\}, \\
\Gamma_5 &= \frac{\left(\Lambda_1 - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1+\|\nabla f_i(w^t) - \Gamma_4\nabla^2 f_i(w^t)\nabla f_i(w^t)\|^2}, \\
\Gamma_6 &= \min\left\{\Gamma_5, \frac{\Lambda_1}{\|\nabla f_i(w^t) - \Gamma_4\nabla^2 f_i(w^t)\nabla f_i(w^t)\|^2}\right\}, \\
\Lambda_1 &= f_i(w^t) - \Gamma_4\left\|\nabla f_i(w^t)\right\|^2 + \frac{1}{2}\Gamma_4^2\left\langle \nabla^2 f_i(w^t)\nabla f_i(w^t), \nabla f_i(w^t)\right\rangle.
\end{aligned}
$$

To solve (29), we consider the following two-step method similar to (27) and (28):

$$
\begin{aligned}
w^{t+1/2}, s^{t+1/2} = \underset{s\geq 0,\, w\in\mathbb{R}^d}{\operatorname{argmin}}\ & \frac{1-\lambda}{2}\Delta_t + \frac{\lambda}{2}s \\
\text{s.t. }\ & q_{i,t}(w^t) + \left\langle \nabla q_{i,t}(w^t), w - w^t\right\rangle \leq s.
\end{aligned} \qquad (107)
$$

$$
\begin{aligned}
w^{t+1}, s^{t+1} = \underset{s\geq 0,\, w\in\mathbb{R}^d}{\operatorname{argmin}}\ & \frac{1-\lambda}{2}\Delta_{t+\frac{1}{2}} + \frac{\lambda}{2}s \\
\text{s.t. }\ & q_{i,t}(w^{t+1/2}) + \left\langle \nabla q_{i,t}(w^{t+1/2}), w - w^{t+1/2}\right\rangle \leq s.
\end{aligned} \qquad (108)
$$

Note that

$$
\frac{1-\lambda}{2}(s-s^t)^2 + \frac{\lambda}{2}s = \frac{1-\lambda}{2}\left(s - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)^2 + \text{constants w.r.t. } w \text{ and } s.
$$

Then, (107) is equivalent to solving

$$w^{t+1/2}, s^{t+1/2} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \|w - w^t\|^2 + \left(s - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)^2 \quad (109)$$
$$\text{s.t. } q_{i,t}(w^t) + \left\langle \nabla q_{i,t}(w^t), w - w^t \right\rangle \leq s.$$

It follows from Lemma C.4 of [cite] that the closed form solution to (107) is

$$w^{t+1/2} = w^t - \min\left\{ \frac{\left(q_{i,t}(w^t) - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1 + \|\nabla q_{i,t}(w^t)\|^2}, \frac{q_{i,t}(w^t)}{\|\nabla q_{i,t}(w^t)\|^2} \right\} \nabla q_{i,t}(w^t),$$
$$s^{t+1/2} = \left(\left(s^t - \frac{\lambda}{2(1-\lambda)}\right) + \frac{\left(q_{i,t}(w^t) - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1 + \|\nabla q_{i,t}(w^t)\|^2} \right)_+,$$

where we denote $(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$.

Note that $q_{i,t}(w^t) = f_i(w^t)$ and $\nabla q_{i,t}(w^t) = \nabla f_i(w^t)$. To simplify the notation, denote

$$\Gamma_3 = \frac{\left(f_i(w^t) - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1 + \|\nabla f_i(w^t)\|^2},$$

and

$$\Gamma_4 = \min\left\{\Gamma_3, \frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}\right\}.$$

Then, we have

$$w^{t+1/2} = w^t - \Gamma_4 \nabla f_i(w^t),$$
$$s^{t+1/2} = \left(\left(s^t - \frac{\lambda}{2(1-\lambda)}\right) + \Gamma_3\right)_+.$$

In a similar way, we can get the closed form solution to (108), which is given as

$$w^{t+1} = w^{t+1/2} - \min\left\{ \frac{\left(q_{i,t}(w^{t+1/2}) - \left(s^{t+1/2} - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1 + \|\nabla q_{i,t}(w^{t+1/2})\|^2}, \frac{q_{i,t}(w^{t+1/2})}{\|\nabla q_{i,t}(w^{t+1/2})\|^2} \right\} \nabla q_{i,t}(w^{t+1/2}),$$
$$s^{t+1} = \left(\left(s^{t+1/2} - \frac{\lambda}{2(1-\lambda)}\right) + \frac{\left(q_{i,t}(w^{t+1/2}) - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1 + \|\nabla q_{i,t}(w^{t+1/2})\|^2} \right)_+.$$

Note that

$$\begin{aligned} q_{i,t}(w^{t+1/2}) &= f_i(w^t) - \left\langle \nabla f_i(w^t), \Gamma_4 \nabla f_i(w^t) \right\rangle + \frac{1}{2} \left\langle \nabla^2 f_i(w^t) \Gamma_4 \nabla f_i(w^t), \Gamma_4 \nabla f_i(w^t) \right\rangle \\ &= f_i(w^t) - \Gamma_4 \left\|\nabla f_i(w^t)\right\|^2 + \frac{1}{2}\Gamma_4^2 \left\langle \nabla^2 f_i(w^t) \nabla f_i(w^t), \nabla f_i(w^t) \right\rangle \\ &\triangleq \Lambda_1 \end{aligned}$$

and

$$\begin{aligned}\nabla q_{i,t}(w^{t+1/2}) &= \nabla f_i(w^t) + \nabla^2 f_i(w^t)(w^{t+1/2} - w^t)\\ &= \nabla f_i(w^t) - \Gamma_4 \nabla^2 f_i(w^t)\nabla f_i(w^t).\end{aligned}$$

Again, to simplify the notation, we denote

$$\begin{aligned}\Gamma_5 &= \frac{\left(q_{i,t}(w^{t+1/2}) - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1 + \|\nabla q_{i,t}(w^{t+1/2})\|^2}\\ &= \frac{\left(\Lambda_1 - \left(s^t - \frac{\lambda}{2(1-\lambda)}\right)\right)_+}{1 + \|\nabla f_i(w^t) - \Gamma_4 \nabla^2 f_i(w^t)\nabla f_i(w^t)\|^2},\end{aligned}$$

and

$$\begin{aligned}\Gamma_6 &= \min\left\{\Gamma_5, \frac{q_{i,t}(w^{t+1/2})}{\|\nabla q_{i,t}(w^{t+1/2})\|^2}\right\}\\ &= \min\left\{\Gamma_5, \frac{\Lambda_1}{\|\nabla f_i(w^t) - \Gamma_4 \nabla^2 f_i(w^t)\nabla f_i(w^t)\|^2}\right\}.\end{aligned}$$

Then, we have

$$\begin{aligned}w^{t+1} &= w^{t+1/2} - \Gamma_6 \nabla q_{i,t}(w^{t+1/2})\\ &= w^t - \Gamma_4 \nabla f_i(w^t) - \Gamma_6\left(\nabla f_i(w^t) - \Gamma_4 \nabla^2 f_i(w^t)\nabla f_i(w^t)\right)\\ &= w^t - (\Gamma_4 + \Gamma_6)\nabla f_i(w^t) + \Gamma_6\Gamma_4 \nabla^2 f_i(w^t)\nabla f_i(w^t),\\ s^{t+1} &= \left(\left(s^{t+1/2} - \frac{\lambda}{2(1-\lambda)}\right) + \Gamma_5\right)_+\\ &= \left(\left(\left(\left(s^t - \frac{\lambda}{2(1-\lambda)}\right) + \Gamma_3\right)_+ - \frac{\lambda}{2(1-\lambda)}\right) + \Gamma_5\right)_+.\end{aligned}$$

F.9. Proof of Lemma 18

The following lemma gives a closed form for two step method `SP2max`$^+$.

Lemma 18 *(`SP2max`$^+$) The w^{t+1} and s^{t+1} update is given by*

$$\begin{aligned}w^{t+1} &= w^t - (\Gamma_1 + \Gamma_3)\nabla f_i(w^t) + \Gamma_3\Gamma_1 \nabla^2 f_i(w^t)\nabla f_i(w^t),\\ s^{t+1} &= \max\left\{\Gamma_2 - \frac{\lambda}{2(1-\lambda)}\left\|\nabla f_i(w^t) - \Gamma_1 \nabla^2 f_i(w^t)\nabla f_i(w^t)\right\|^2, 0\right\},\end{aligned}$$

where

$$\begin{aligned}\Gamma_1 &= \min\left\{\frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}, \frac{\lambda}{2(1-\lambda)}\right\},\\ \Gamma_2 &= f_i(w^t) - \Gamma_1\left\|\nabla f_i(w^t)\right\|^2 + \tfrac{1}{2}\Gamma_1^2\left\langle \nabla^2 f_i(w^t)\nabla f_i(w^t), \nabla f_i(w^t)\right\rangle,\\ \Gamma_3 &= \min\left\{\frac{\Gamma_2}{\|\nabla f_i(w^t) - \Gamma_1 \nabla^2 f_i(w^t)\nabla f_i(w^t)\|^2}, \frac{\lambda}{2(1-\lambda)}\right\}.\end{aligned}$$

To solve (30), we again consider a two step method similar to (107) and (108):

$$w^{t+1/2}, s^{t+1/2} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1-\lambda}{2}\left\|w-w^t\right\|^2 + \frac{\lambda}{2}s \quad \text{s.t. } q_{i,t}(w^t) + \left\langle \nabla q_{i,t}(w^t), w-w^t\right\rangle \leq s. \tag{110}$$

$$w^{t+1}, s^{t+1} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1-\lambda}{2}\left\|w-w^{t+1/2}\right\|^2 + \frac{\lambda}{2}s \quad \text{s.t. } q_{i,t}(w^{t+1/2}) + \left\langle \nabla q_{i,t}(w^{t+1/2}), w-w^{t+1/2}\right\rangle \leq s. \tag{111}$$

Note that (110) is equivalent to solving

$$w^{t+1/2}, s^{t+1/2} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2}\left\|w-w^t\right\|^2 + \frac{\lambda}{2(1-\lambda)}s \quad \text{s.t. } q_{i,t}(w^t) + \left\langle \nabla q_{i,t}(w^t), w-w^t\right\rangle \leq s. \tag{112}$$

It follows from Lemma 3.1 in [cite] that the closed form solution to (112) is

$$\begin{aligned}
w^{t+1/2} &= w^t - \min\left\{\frac{q_{i,t}(w^t)}{\|\nabla q_{i,t}(w^t)\|^2}, \frac{\lambda}{2(1-\lambda)}\right\}\nabla q_{i,t}(w^t) \\
&= w^t - \min\left\{\frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}, \frac{\lambda}{2(1-\lambda)}\right\}\nabla f_i(w^t) \\
&= w_t - \Sigma_1 \nabla f_i(w^t), \\
s^{t+1/2} &= \max\left\{q_{i,t}(w^t) - \frac{\lambda}{2(1-\lambda)}\left\|\nabla q_{i,t}(w^t)\right\|^2, 0\right\} \\
&= \max\left\{f_i(w^t) - \frac{\lambda}{2(1-\lambda)}\left\|\nabla f_i(w^t)\right\|^2, 0\right\},
\end{aligned}$$

where we denote

$$\Gamma_1 = \min\left\{\frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2}, \frac{\lambda}{2(1-\lambda)}\right\}.$$

Note that

$$\begin{aligned}
q_{i,t}(w^{t+1/2}) &= f_i(w^t) - \Gamma_1\left\|\nabla f_i(w^t)\right\|^2 + \tfrac{1}{2}\Gamma_1^2\left\langle \nabla^2 f_i(w^t)\nabla f_i(w^t), \nabla f_i(w^t)\right\rangle \\
&:= \Gamma_2,
\end{aligned}$$

and

$$\begin{aligned}
q_{i,t}(w^{t+1/2}) &= \nabla f_i(w^t) + \nabla^2 f_i(w^t)(w^{t+1/2} - w^t) \\
&= \nabla f_i(w^t) - \Gamma_1 \nabla^2 f_i(w^t)\nabla f_i(w^t).
\end{aligned}$$

Similarly, we have the closed form solution to (111) given as

$$\begin{aligned}
w^{t+1} &= w^{t+1/2} - \min\left\{\frac{q_{i,t}(w^{t+1/2})}{\left\|\nabla q_{i,t}(w^{t+1/2})\right\|^2}, \frac{\lambda}{2(1-\lambda)}\right\}\nabla q_{i,t}(w^{t+1/2}) \\
&= w^{t+1/2} - \min\left\{\frac{\Gamma_2}{\|\nabla f_i(w^t) - \Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t)\|^2}, \frac{\lambda}{2(1-\lambda)}\right\}\left(\nabla f_i(w^t) - \Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t)\right) \\
&= w^{t+1/2} - \Gamma_3\left(\nabla f_i(w^t) - \Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t)\right) \\
&= w^t - (\Gamma_1 + \Gamma_3)\nabla f_i(w^t) + \Gamma_3\Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t) \\
s^{t+1} &= \max\left\{q_{i,t}(w^{t+1/2}) - \frac{\lambda}{2(1-\lambda)}\left\|\nabla q_{i,t}(w^{t+1/2})\right\|^2, 0\right\} \\
&= \max\left\{\Gamma_2 - \frac{\lambda}{2(1-\lambda)}\left\|\nabla f_i(w^t) - \Gamma_1\nabla^2 f_i(w^t)\nabla f_i(w^t)\right\|^2, 0\right\}.
\end{aligned}$$

F.10. Proof of Lemma 7

In GLMs, the unregularized problem (30) becomes

$$\begin{aligned}
w^{t+1}, s^{t+1} = \underset{s\geq 0,\ w\in\mathbb{R}^d}{\operatorname{argmin}} & \frac{1}{2}\left\|w - w^t\right\|^2 + \widetilde{\lambda}s \\
\text{s.t. } & f_i + \left\langle a_i x_i, w - w^t\right\rangle + \tfrac{1}{2}\left\langle h_i x_i x_i^\top (w - w^t), w - w^t\right\rangle \leq s,
\end{aligned} \tag{113}$$

where $\widetilde{\lambda} \coloneqq \frac{\lambda}{2(1-\lambda)}$, and we denote $f_i = f_i(w^t)$, $a_i \coloneqq \phi_i'(x_i^\top w - y_i)$, $h_i \coloneqq \phi_i''(x_i^\top w - y_i)$ for short.

Denote $\triangle \coloneqq w - w^t$. Then, problem (113) reduces to

$$\begin{aligned}
\min_{s\geq 0,\ \triangle\in\mathbb{R}^d} & \frac{1}{2}\|\triangle\|^2 + \widetilde{\lambda}s \\
\text{s.t. } & f_i + a_i x_i^\top\triangle + \tfrac{1}{2}h_i\triangle^\top x_i x_i^\top\triangle \leq s.
\end{aligned} \tag{114}$$

Note that we want to minimize $\|\triangle\|^2$. Together with the above constraint, we can conclude that $\triangle$ must be a multiple of x_i since any other component will not help satisfy the constraint but increase $\|\triangle\|^2$. Let $\triangle = cx_i$ and $\ell = \|x_i\|^2$, then problem (114) becomes

$$\begin{aligned}
\min_{s\geq 0,\ c\in\mathbb{R}} & \frac{1}{2}c^2\ell + \widetilde{\lambda}s \\
\text{s.t. } & f_i + a_i\ell c + \tfrac{1}{2}h_i\ell^2 c^2 \leq s.
\end{aligned} \tag{115}$$

The corresponding Lagrangian function is then given as

$$L(s, c, \nu_1, \nu_2) = \frac{1}{2}c^2\ell + \widetilde{\lambda}s + \nu_1(f_i + a_i\ell c + \tfrac{1}{2}h_i\ell^2c^2 - s) - \nu_2 s,$$

where $\nu_1, \nu_2 \geq 0$ are the Lagrangian multipliers. The KKT conditions are thus

$$\begin{aligned} f_i + a_i \ell c + \tfrac{1}{2} h_i \ell^2 c^2 - s \leq 0, \\ s \geq 0, \quad \nu_1 \geq 0, \quad \nu_2 \geq 0, \\ \nu_1 (f_i + a_i \ell c + \tfrac{1}{2} h_i \ell^2 c^2 - s) = 0, \quad \nu_2 s = 0, \\ \widetilde{\lambda} - \nu_1 - \nu_2 = 0, \\ \ell c + \nu_1 a_i \ell + \nu_1 h_i \ell^2 c = 0. \end{aligned}$$

By checking the complementary conditions, the solution to the above KKT equations has three cases, which are summarized below.

Case I: The Lagragian multiplier $\nu_2 = 0$. In which case $\nu_1^\star = \widetilde{\lambda}$, $\nu_2^\star = 0$, $c^\star = -\frac{\widetilde{\lambda} a_i}{1 + \widetilde{\lambda} h_i \ell}$, and

$$s^\star = f_i + a_i \ell c^\star + \frac{1}{2} h_i \ell^2 c^{\star 2} = f_i - \frac{\widetilde{\lambda} a_i^2 \ell}{1 + \widetilde{\lambda} h_i \ell} + \frac{h_i \widetilde{\lambda}^2 a_i^2 \ell^2}{2(1 + \widetilde{\lambda} h_i \ell)^2},$$

which is feasible if $s^\star \geq 0$. The resulting objective function is $\frac{1}{2} c^{\star 2} \ell + \widetilde{\lambda} s^\star$, which is ≥ 0.

Case II: The Lagragian multiplier $\nu_1 = 0$. In which case $\nu_1^\star = 0$, $\nu_2^\star = \widetilde{\lambda}$, $c^\star = 0$, $s^\star = 0$, which is feasible if $f_i = 0$. The objective function is 0 in this case and the variable w is unchanged since $w - w^t = c^\star x_i = 0$.

Case III: Neither Lagragian multiplier is zero. In which case there are two possible solutions for c given by $c^\star = \frac{-a_i \pm \sqrt{a_i^2 - 2 h_i f_i}}{h_i \ell}$, $\nu_1^\star = -\frac{c}{a_i + h_i \ell c}$, $\nu_2^\star = \widetilde{\lambda} + \frac{c}{a_i + h_i \ell c}$, $s^\star = 0$. Note that

$$a_i + h_i \ell c = \pm \sqrt{a_i^2 - 2 h_i f_i}.$$

Consequently to guarantee that the Lagrangian multipliers ν_1 and ν_2 are non-negative, we must have $c^\star = \frac{-a_i + \sqrt{a_i^2 - 2 h_i f_i}}{h_i \ell}$ and in this case the objective function equals $\frac{1}{2} c^{\star 2} \ell \geq 0$.

As a summary, if $f_i = 0$, Case II is the optimal solution. Alternatively if $f_i > 0$ and if

$$\widetilde{s} = f_i - \frac{\widetilde{\lambda} a_i^2 \ell}{1 + \widetilde{\lambda} h_i \ell} + \frac{h_i \widetilde{\lambda}^2 a_i^2 \ell^2}{2(1 + \widetilde{\lambda} h_i \ell)^2}$$

is non-negative then Case I is the optimal solution. Otherwise, Case III with $c^\star = \frac{-a_i + \sqrt{a_i^2 - 2 h_i f_i}}{h_i \ell}$ is the optimal solution.

Therefore, the optimal solution to (113) is then

$$\begin{aligned} w^{t+1} &= w^t + c^\star x_i, \\ s^{t+1} &= \max\{\widetilde{s},\ 0\}, \end{aligned}$$

where

$$c^\star = \begin{cases} 0, & \text{if } f_i = 0 \\ -\frac{\widetilde{\lambda} a_i}{1 + \widetilde{\lambda} h_i \ell}, & \text{if } f_i > 0 \text{ and } \widetilde{s} \geq 0, \\ \frac{-a_i + \sqrt{a_i^2 - 2 h_i f_i}}{h_i \ell}, & \text{otherwise.} \end{cases}$$

Appendix G. Additional Numerical Experiments

G.1. Non-convex problems

For the non-convex experiements, we used the Python Package `pybenchfunction` available on github `Python_Benchmark_Test_Optimization_Function_Single_Objective`.

G.1.1. PermDβ^+ is an incorrect implementation of PermDβ.

We note here that the PermDβ^+ implemented is given by

$$\texttt{PermD}\beta^+(x) := \sum_{i=1}^{d}\sum_{j=1}^{d}\left((j^i+\beta)\left(\left(\frac{x_j}{j}\right)^i - 1\right)\right)^2.$$

which is different than the standard $\texttt{PermD}\beta$ function which is given by

$$\texttt{PermD}\beta(x) := \sum_{i=1}^{d}\left(\sum_{j=1}^{d}(j^i+\beta)\left(\left(\frac{x_j}{j}\right)^i - 1\right)\right)^2.$$

We believe this is a small mistake, which is why we have introduced the plus in $\texttt{PermD}\beta^+$ to distinguish this function from the standard $\texttt{PermD}\beta$ function. Yet, the $\texttt{PermD}\beta^+$ is still an interesting non-convex problem, and thus we have used it in our experiments despite this small alteration.

G.1.2. The Levy N. 13 and Rosenbrock problems

Here we provide two additional experiments on the non-convex function Levy N. 13 and Rosenbrock that complement the findings in 3.1.

For the Levy N. 13 function in Figure 1 we have that again `SP2` converges in 10 epochs to the global minima. In contrast Newton's method converges immediately to a local maxima, that can be easily seen on the surface plot of the right of Figure 1.

The one problem where `SP2` was not the fastest was on the Rosenbrock function, see Figure 4. Here `Newton` was the fastest, converging in under 10 epochs. But note, this problem was designed to emphasize the advantages of Newton over gradient descent.

G.2. Matrix Completion

Assume a set of known values $\{a_{i,j}\}_{(i,j)\in\Omega}$ where Ω is a set of known elements of the matrix, and we want to determine the missing elements. One approach is solving the *matrix completion* problem

$$\min_{U,V}\sum_{(i,j)\in\Omega}\frac{1}{2}(u_i^Tv_j - a_{i,j})^2, \tag{116}$$

where $U=[u_i]_{(i,j)\in\Omega}$ and $V=[v_j]_{(i,j)\in\Omega}$. With the solution to (116), we then use $U^\top V$ as an approximation to the complete matrix $A=[a_{i,j}]_{i,j=1,\ldots,n}$.

Despite (116) being a non-convex problem, if there exists an *interpolating* solution to (116), that is one where $u_i^Tv_j = a_{i,j},$ for $(i,j)\in\Omega$, then the SP2 method can solve (116).

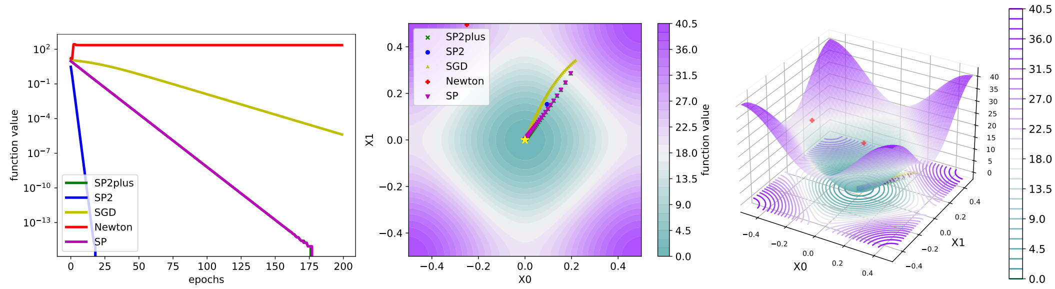


Figure 2: The Rastrigin function where Left: we plot $f(x)$ across epochs Middle: level set plot, Right: Surface plot. SP2 is in blue , SP2^{+} is in green , SGD is in yellow and Newton is in red.

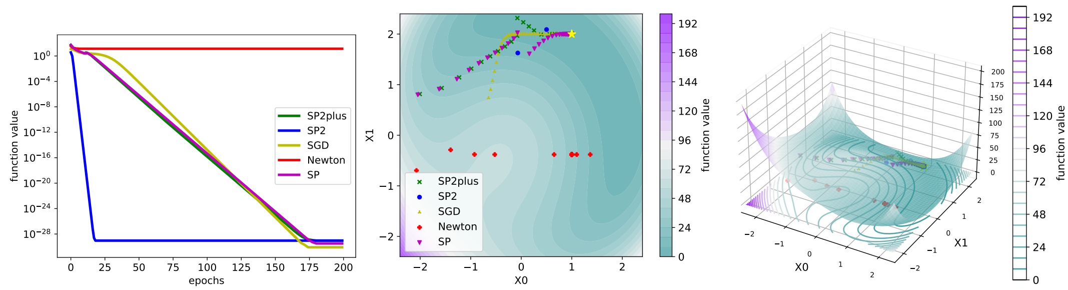


Figure 3: The PermDβ^{+} function with $\beta = 0.5$ where Left: we plot $f(x)$ across epochs Middle: level set plot, Right: Surface plot. SP2 is in blue , SP2^{+} is in green , SGD is in yellow and Newton in red.

Indeed, the SP2 can be applied to (116) by sampling a single pair $(i, j) \in \Omega$ uniformly, then projecting onto the quadratic

$$u_i^{k+1}, v_j^{k+1} = \operatorname{argmin}_{u,v} \frac{1}{2} \left\| u - u_i^k \right\|^2 + \frac{1}{2} \left\| v - v_j^k \right\|^2 \text{ subject to } u^\top v = a_{i,j}. \tag{117}$$

This projection can be solved as we detail in Theorem 10 in Appendix E.

We compared our method 117 to a specialized variant of SGD for online matrix completion described in Jin et al. (2016), see Figure 5. To compare the two methods we generated a rank $k = 2$ matrix $A \in \mathbb{R}^{100 \times 50}$. We selected a subset entries with probability $p = 0.1, 0.2$ or 0.3 to form our set Ω_{init} that was used to obtain an initial estimate U_0, V_0 using rank-k SVD method as described in Jin et al. (2016). We extensively tuned the step size of SGD using a grid search, and the method labelled Non-convex SGD is the resulting run of SGD with the best step size. We also show how sensitive SGD is to this step size, by including the run of SGD with step sizes that were only a factor of 2 to 4 away from the optimal, which greatly degrades the performance of SGD. In contrast, SP2 worked with no tuning, and matches the performance of SGD with the optimal step size in the $p = 0.1$ experiment, and outperforms SGD in the experiments with more measurements as can be seen in the $p = 0.2$ and $p = 0.3$ figures.

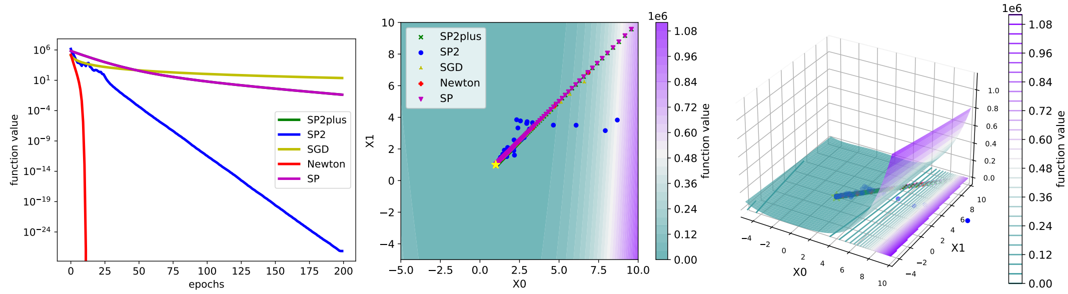


Figure 4: The Rosenbrock function where Left: we plot $f(x)$ across epochs Middle: level set plot, Right: Surface plot. SP2 is in blue , SP2^+ is in green , SGD is in yellow and Newton is in red .

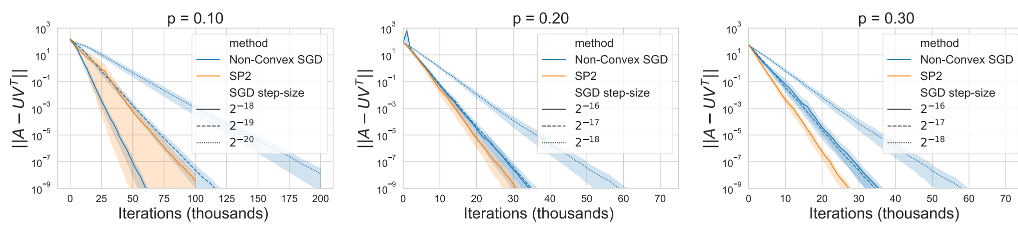


Figure 5: Recovery error for matrix completion. Left: Using 10%, Middle: using 20% and Right: using 30% of the entries of A to form Ω, respectively. Shaded region corresponds to 5 repeated runs.

G.3. Convex classification

In this experiment, we test the proposed methods on a logistic regression problem and compare them with some state-of-the-art methods (e.g., SGD, SP, and ADAM). In particular, we consider the problem of minimizing the following loss function $f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w) + \frac{\sigma}{2}\|w\|_2^2$, where $f_i(w) = \phi_i(x_i^\top w)$ with $\phi_i(t) = \ln(1 + e^{-y_i t})$. Here, $\{(x_i, y_i) \in \mathbb{R}^{d+1}\}_{i=1}^n$ stands for the feature-label pairs and $\sigma > 0$ is the regularization parameter. We can control how far each problem is from interpolation by increasing σ. When $\sigma > 0$ the problem cannot interpolate, and thus we expect to see a benefit of the slack methods in Section C over SP2^+.

We used two data sets: colon-cancer Alon et al. (1999) and mushrooms West et al. (2001), both of which interpolate when $\sigma = 0$.

We compare the proposed methods SP2^+ (18), SP2L2^+ (Lemma 15), SP2L1^+ (Lemma 17), and SP2max^+ (Lemma 18) with SGD, SP (5), and ADAM on both data sets with three regularizations $\sigma \in \{0,\ 0.001,\ 0.008\}$ and with momentum set to 0.3. For the SGD method, we use a learning rate $L_{\max}/\sqrt{t}$ in the t-th iteration, where $L_{\max} = \frac{1}{4}\max_{i=1,\ldots,n}\|x_i\|^2$ denotes the smoothness constant of the loss function. We chose λ for SP2L2^+, SP2L1^+, and SP2max^+ using a grid search of $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$, the details are in Section G.

The gradient norm and loss evaluated at each epoch are presented in Figures 6 and 13 (see Appendix G). We see that SP2 methods converge much faster than classical methods

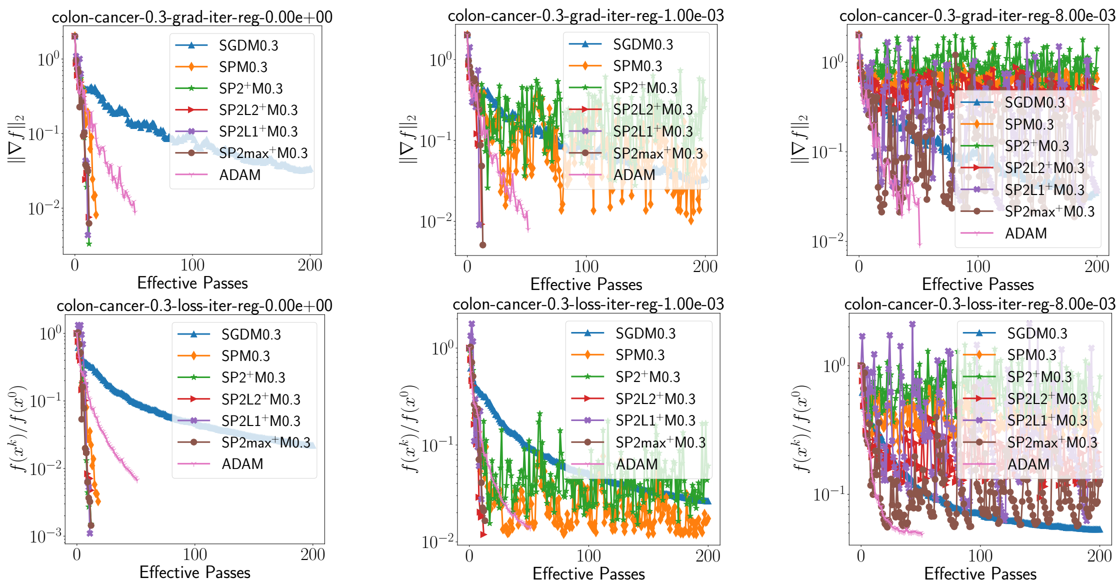


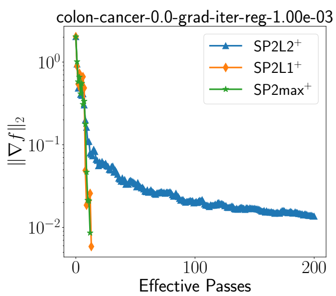
Figure 6: Colon-cancer: gradient norm and loss at each epoch with momentum being 0.3. Left: $\sigma = 0$, Middle: $\sigma = 0.001$ and Right $\sigma = 0.008$.

(e.g., `SGD`, `SP`, `ADAM`) and need fewer epochs to achieve the tolerance when σ is small (left and middle plots). However, they can all fail when the problem is far from interpolation, e.g., when $\sigma = 8 \times 10^{-3}$. The running time used for each algorithm to achieve either the tolerance or maximum number of epochs for both data sets is presented in Figure 14 (see Appendix G).

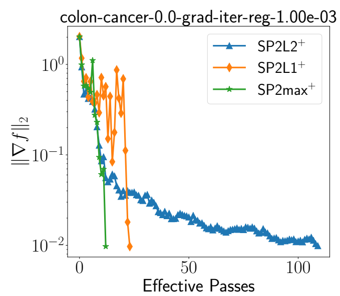
G.4. Additional Convex Experiments

We set the desired tolerance for each algorithm to 0.01, and set the maximum number of epochs for each algorithm to 200 in colon-cancer and 30 in mushrooms. To choose an optimal slack parameter λ for $\texttt{SP2L2}^+$, $\texttt{SP2L1}^+$, and $\texttt{SP2max}^+$, we test these three methods on a uniform grid $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$ with $\sigma = 0.001$. The gradient norm and loss evaluated at each epoch are presented in Figures 7-10 (see Appendix G). It can be seen that $\texttt{SP2L2}^+$ performs best when $\lambda = 0.9$ in colon-cancer and $\lambda = 0.1$ in mushrooms, $\texttt{SP2L1}^+$ and $\texttt{SP2max}^+$ perform best when $\lambda = 0.1$ in both data sets. Therefore, we set $\lambda = 0.9$ for $\texttt{SP2L2}^+$ in colon-cancer and fix $\lambda = 0.1$ in other cases.

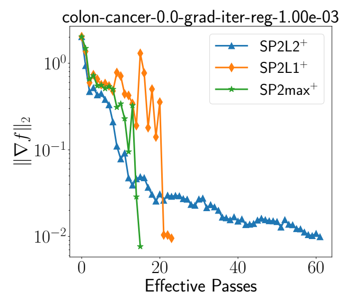
Under the same setting as in Section G.3, we also compare the `SP2max` and $\texttt{SP2max}^+$ methods on a grid $\lambda = [0.001\ 0.01 : 0.01 : 0.05]$ with $\sigma = 0$. The gradient norm and loss evaluated at each epoch are presented in Figures 11-12 (see Appendix G). As we observe, $\texttt{SP2max}^+$ always outperforms the `SP2max` method.



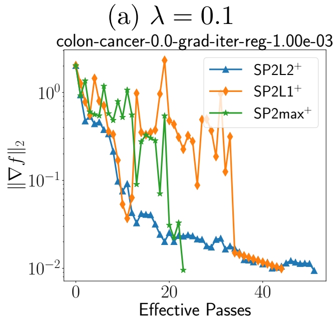
(a) $\lambda = 0.1$



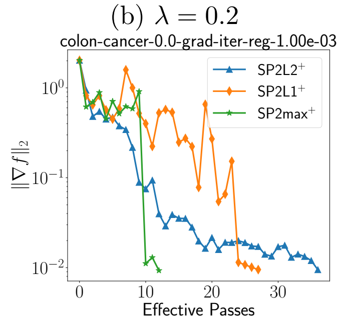
(b) $\lambda = 0.2$



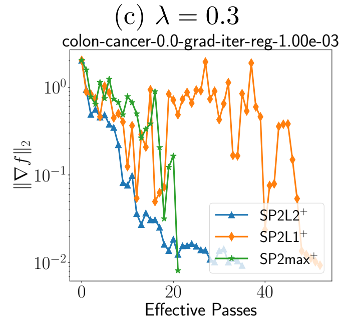
(c) $\lambda = 0.3$



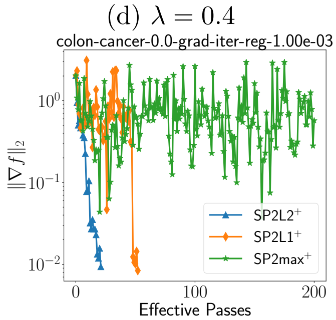
(d) $\lambda = 0.4$



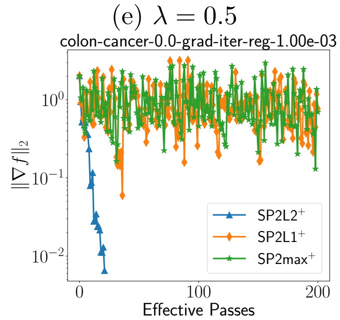
(e) $\lambda = 0.5$



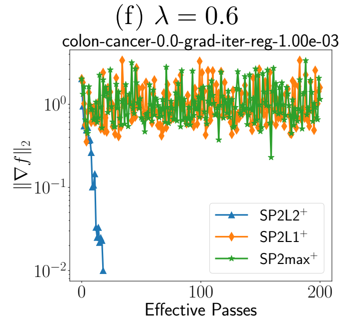
(f) $\lambda = 0.6$



(g) $\lambda = 0.7$

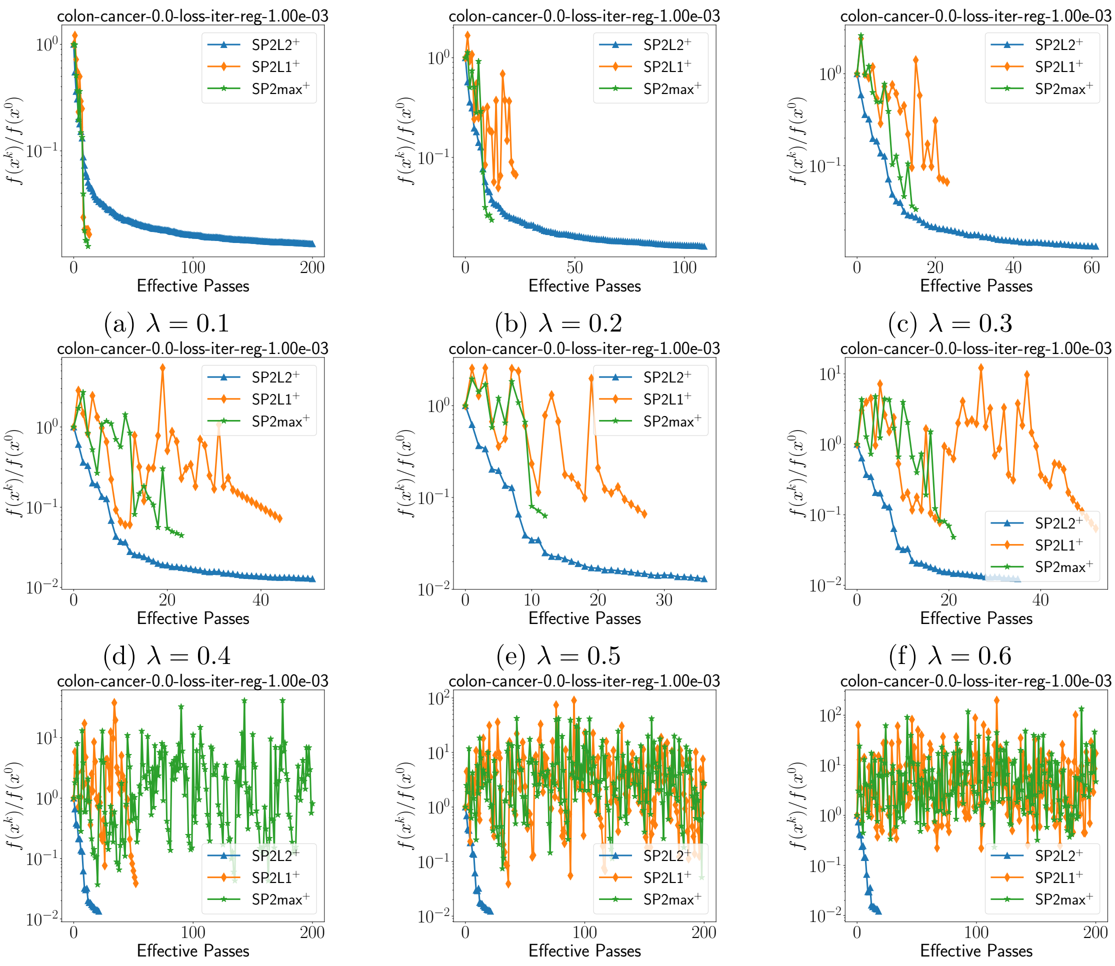


(h) $\lambda = 0.8$



(i) $\lambda = 0.9$

Figure 7: Colon-cancer: gradient norm at each epoch with different λ.

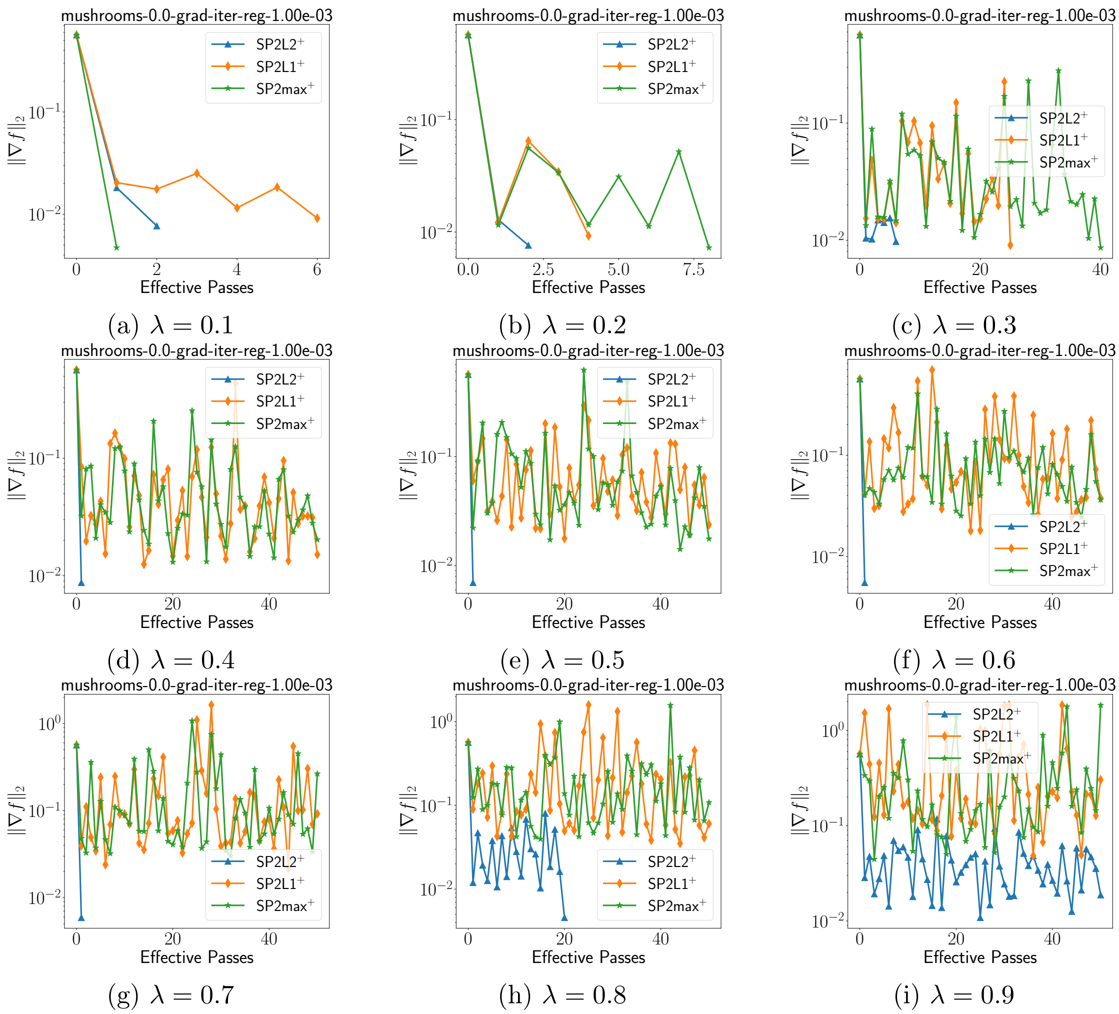


(a) $\lambda = 0.1$ (b) $\lambda = 0.2$ (c) $\lambda = 0.3$

(d) $\lambda = 0.4$ (e) $\lambda = 0.5$ (f) $\lambda = 0.6$

(g) $\lambda = 0.7$ (h) $\lambda = 0.8$ (i) $\lambda = 0.9$

Figure 8: Colon-cancer: loss at each epoch with different λ.

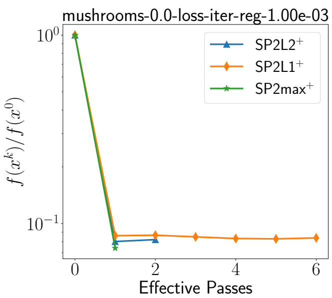


(a) $\lambda = 0.1$ (b) $\lambda = 0.2$ (c) $\lambda = 0.3$

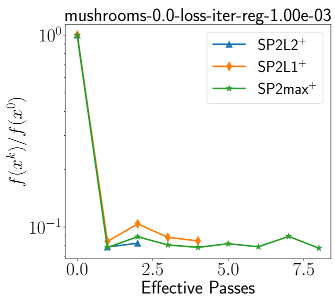
(d) $\lambda = 0.4$ (e) $\lambda = 0.5$ (f) $\lambda = 0.6$

(g) $\lambda = 0.7$ (h) $\lambda = 0.8$ (i) $\lambda = 0.9$

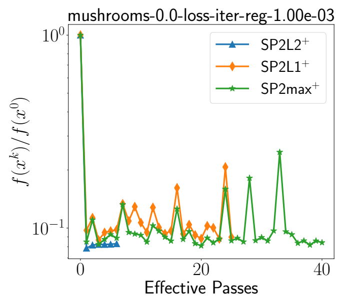
Figure 9: Mushrooms: gradient norm at each epoch with different λ.



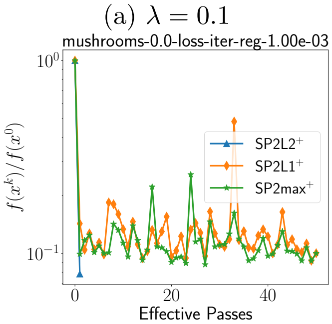
(a) $\lambda = 0.1$



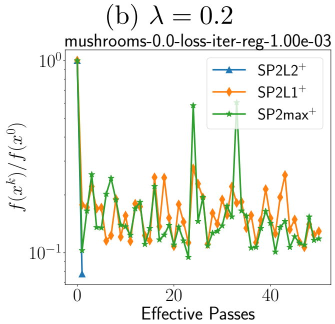
(b) $\lambda = 0.2$



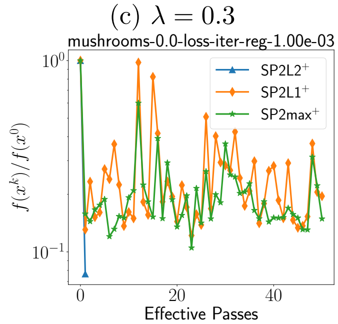
(c) $\lambda = 0.3$



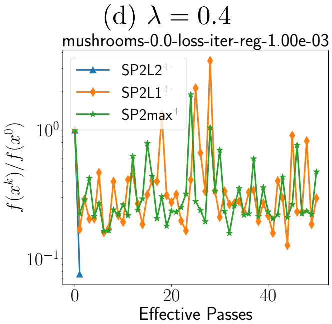
(d) $\lambda = 0.4$



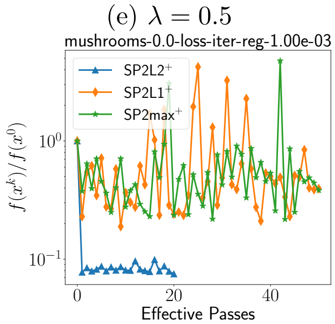
(e) $\lambda = 0.5$



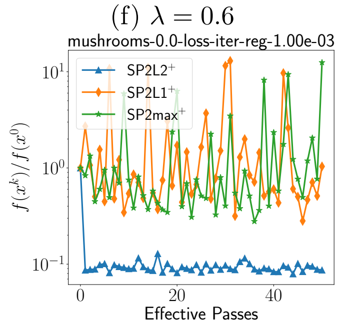
(f) $\lambda = 0.6$



(g) $\lambda = 0.7$

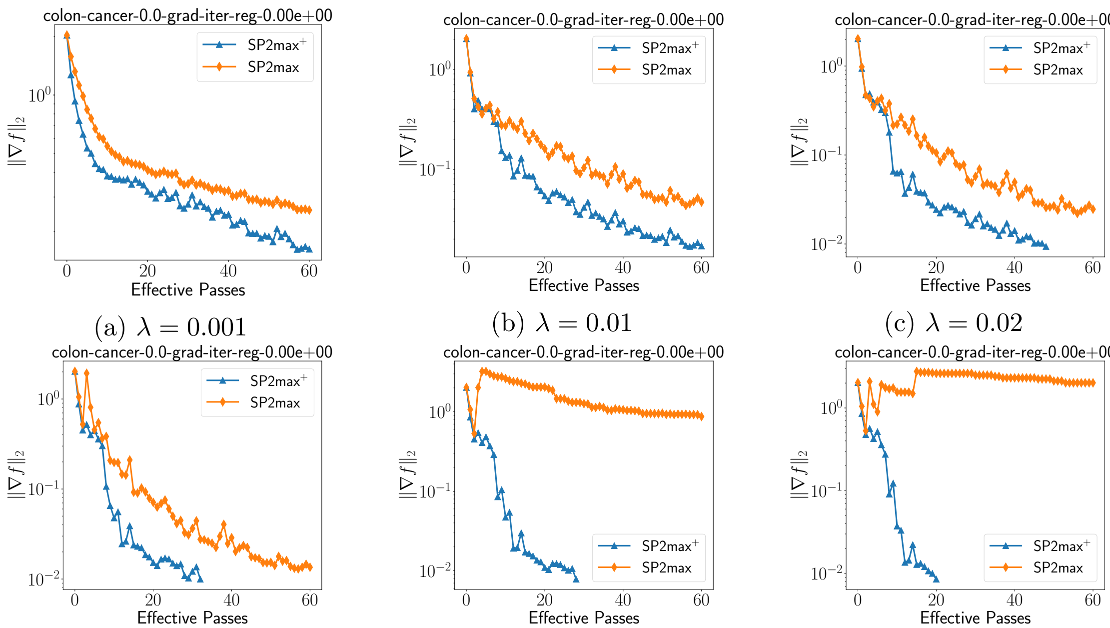


(h) $\lambda = 0.8$



(i) $\lambda = 0.9$

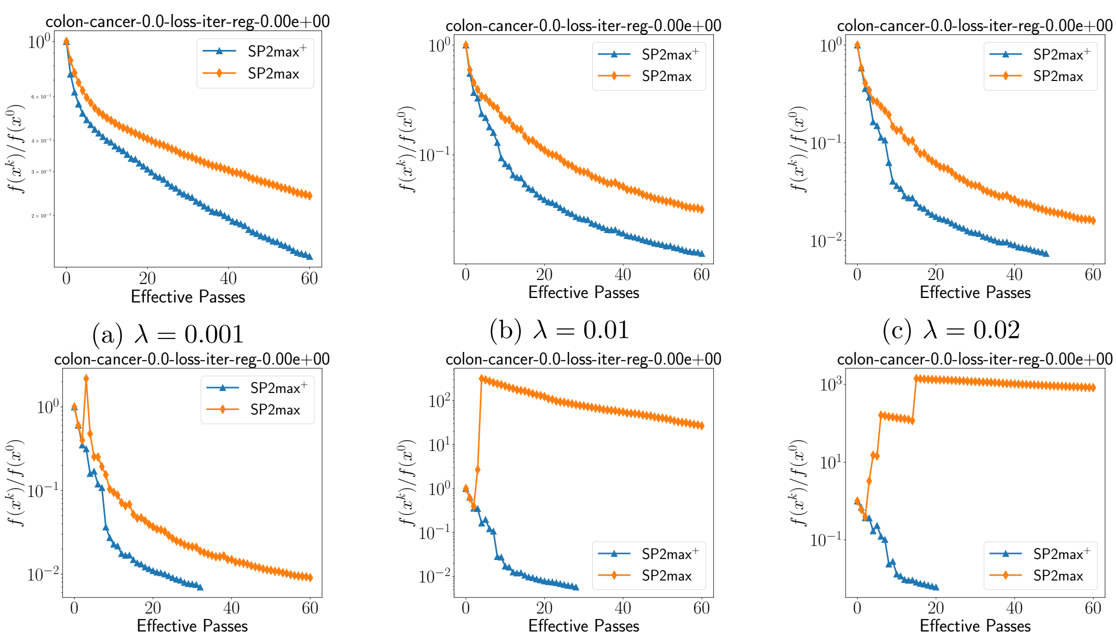
Figure 10: Mushrooms: loss at each epoch with different λ.



(a) $\lambda = 0.001$ (b) $\lambda = 0.01$ (c) $\lambda = 0.02$

(d) $\lambda = 0.03$ (e) $\lambda = 0.04$ (f) $\lambda = 0.05$

Figure 11: Colon-cancer: gradient norm at each epoch with different λ.



(a) $\lambda = 0.001$ (b) $\lambda = 0.01$ (c) $\lambda = 0.02$

(d) $\lambda = 0.03$ (e) $\lambda = 0.04$ (f) $\lambda = 0.05$

Figure 12: Colon-cancer: loss at each epoch with different λ.

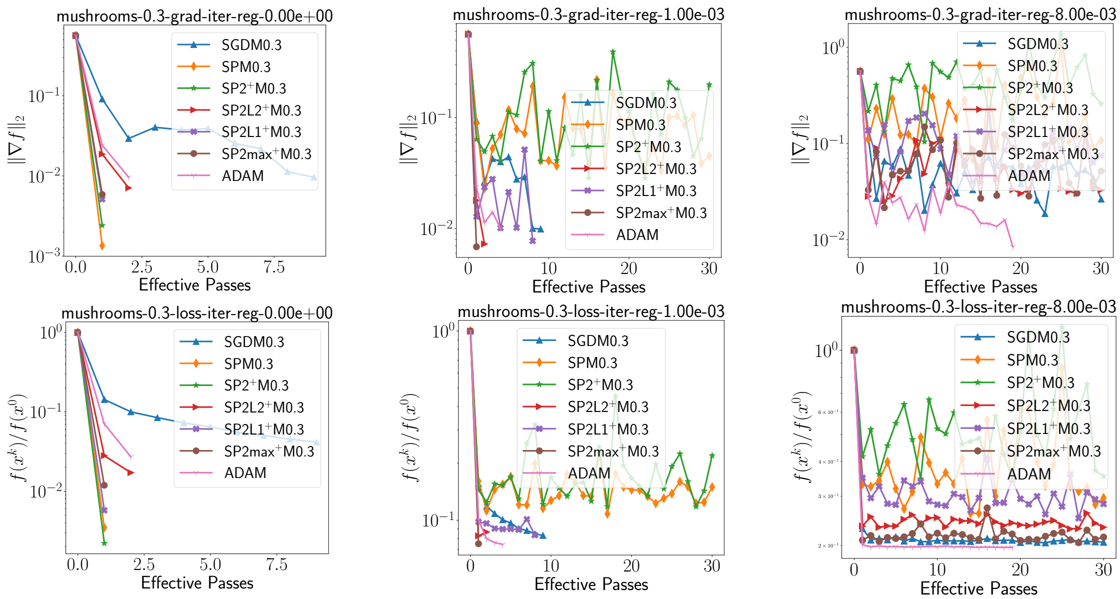


Figure 13: Mushrooms: gradient norm and loss at each epoch with momentum being 0.3.

Figure 14: Running time in seconds.