Adaptive Proximal Gradient Methods for Structured Neural Networks

Jihun Yun
KAIST
arcprime@kaist.ac.kr

Aurélie C. Lozano
IBM T.J. Watson Research Center
aclozano@us.ibm.com

Eunho Yang
KAIST, AITRICS
eunhoy@kaist.ac.kr

Abstract

We consider the training of structured neural networks where the regularizer can be non-smooth and possibly non-convex. While popular machine learning libraries have resorted to stochastic (adaptive) subgradient approaches, the use of proximal gradient methods in the stochastic setting has been little explored and warrants further study, in particular regarding the incorporation of adaptivity. Towards this goal, we present a general framework of stochastic proximal gradient descent methods that allows for arbitrary positive preconditioners and lower semi-continuous regularizers. We derive two important instances of our framework: (i) the first proximal version of ADAM, one of the most popular adaptive SGD algorithm, and (ii) a revised version of PROXQUANT [1] for quantization-specific regularizers, which improves upon the original approach by incorporating the effect of preconditioners in the proximal mapping computations. We provide convergence guarantees for our framework and show that adaptive gradient methods can have faster convergence in terms of constant than vanilla SGD for sparse data. Lastly, we demonstrate the superiority of stochastic proximal methods compared to subgradient-based approaches via extensive experiments. Interestingly, our results indicate that the benefit of proximal approaches over sub-gradient counterparts is more pronounced for non-convex regularizers than for convex ones.

1 Introduction

We study the regularized training of neural networks, which can be formulated as the following (stochastic) optimization problem

$$\underset{\theta\in\mathbb{R}^d}{\text{minimize}}\ F(\theta) := \overbrace{\mathbb{E}_{\xi\sim\mathbb{P}}\big[f(\theta;\xi)\big]}^{f(\theta)} + \mathcal{R}(\theta) \tag{1}$$

where $\theta \in \mathbb{R}^d$ represents the network parameter, ξ is the random variable representing mini-batch data samples, and $\mathcal{R}(\cdot)$ is a regularizer encouraging low-dimensional structural constraints on θ.

The technique of regularization is ubiquitous in machine learning as it can effectively prevent overfitting and yield better generalization. The ℓ_1-regularized training for Lasso estimators/sparse Gaussian graphical model (GMRF) estimation [2, 3] and ℓ_2 weight decay [4] on parameters are prototypical examples. In the context of deep learning, important instances include network pruning [5, 6], which induces a sparse network structure, and network quantization [7, 8, 1], which gives hard constraints so that parameters have only discrete values.

For the *unregularized* case, i.e., when $\mathcal{R}(\theta) = 0$, stochastic gradient descent (SGD) has been a prevalent approach to solve the optimization problem stated in (1). At each iteration, SGD evaluates the gradient on a randomly chosen subset of training samples (mini-batch). While vanilla SGD employs a uniform learning rate for all coordinates, several adaptive variants have been proposed to

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Table 1: Comparison among *stochastic* (or *online*) PGD for solving the problem in (1).

Algorithm	Non-convex Loss	Non-convex Regularizer	Arbitrary Preconditioner	Momentum	Convergence Guarantee
ADAGRAD [9]	✗	✗	△ (ADAGRAD)	✗	✓
[10]	✓	✗	✓	✗	✓
[11]	✓	✗	✗	✓	✓
[12]	✓	✗	✗	✓	✓
[13]	✓	✗	✗	✓	✓
[14]	✓	✓	✗	✗	✓
[15]	✓	✓	△ (ADAGRAD)	✗	✓
Prox-SGD [16]	✓	✗	✓	✓	✗
[17]	✓	✓	✗	✓	✓
PROXGEN **(Ours)**	✓	✓	✓	✓	✓

dynamically take advantage of the data geometry by scaling the learning rate for each coordinate by its gradient history. Prime examples of such approaches include ADAGRAD [9], which adjusts the learning rate by the sum of all the past squared gradients, and exponential moving average (EMA) approaches such as RMSPROP [18] and ADAM [19], which scale down the gradients by square roots of exponential moving averages of squared past gradients to essentially limit the scope of the adaptation to only a few recent gradients. In terms of theory, convergence analyses of these unregularized SGD methods, whether adaptive or not, have been well studied both for convex [19, 20] and non-convex [21, 22] loss f cases.

For the *regularized* case, since the regularizer is often *non-smooth* around some region (e.g. the ℓ_1 norm), modern machine learning libraries such as TensorFlow [23] and PyTorch [24] therefore resort to using the *subgradient* of the objective function $F(\theta)$ in (1). Such a strategy is problematic as it may slow down convergence and result in oscillations.

A simple idea to bypass the non-smoothness of a regularizer is via its proximal operator. This idea is the basis of proximal gradient descent (PGD) methods, which first update the parameter using the gradient of the loss function $f(\theta)$ and then perform a proximal mapping of $\mathcal{R}(\theta)$. In the *non-stochastic* case, PGD with both convex and non-convex regularizers has been extensively studied in the literature [25, 26, 11, 12, 27]. Another work, VMFB [28], analyzes the preconditioned gradient descent on convex regularized problems with non-convex loss but does not consider the first-order momentum. In contrast, PGD in the *stochastic* setting has been little explored. [9, 10] consider PGD to solve the stochastic objectives with convex regularizers. Recently, [15] studies non-convex and non-smooth regularized problems for DC (difference of convex) functions and [14, 17] present non-asymptotic analysis for non-convex smooth loss and non-convex regularizers, which is the most general setting, but do not consider the preconditioner in the update rule.

All the aforementioned studies of the stochastic case, however, focus either on limited settings (e.g. [9] only covers the update rule of ADAGRAD) with convex regularizers only, or on pure vanilla gradient descent for non-convex regularizers. Hence, they cannot accommodate all advanced modern optimization algorithms with *preconditioners*, such as adaptive gradient methods. The only exception is PROX-SGD [16], with the caveat that PROX-SGD update rule is *not a pure* PGD. Moreover, the theory in [16] only guarantees convergence, *not how fast* Prox-SGD converges, and the analysis is performed *without* considering preconditioners.

In this paper, we propose an exact framework for stochastic proximal gradient methods with arbitrary positive preconditioners and lower semi-continuous (possibly non-convex) regularizers. With our framework, our goal is to provide theoretical and empirical understanding of stochastic proximal gradient methods for training structured neural networks. Our main contributions can be summarized as follows:

- We propose the first general family of stochastic proximal gradient methods, which we term PROXGEN. We introduce two important instances stemming from our approach: (i) the first proximal version of ADAM [19] and (ii) a revised version of PROXQUANT [1] that improves upon the original approach for quantization-specific regularizers by incorporating the effect of preconditioners when computing proximal mappings.

Algorithm 1 PROXGEN: A **Gen**eral Stochastic **Prox**imal Gradient Method

1: **Input:** Stepsize α_t, $\{\rho_t\}_{t=1}^{t=T} \in [0, 1)$, regularization parameter λ, and small constant $0 < \delta \ll 1$.
2: **Initialize:** $\theta_1 \in \mathbb{R}^d$, $m_0 = 0 \in \mathbb{R}^d$, and $C_0 = O \in \mathbb{R}^{d \times d}$.
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Draw a minibatch sample ξ_t from $\mathbb{P}$
5: $g_t \leftarrow \nabla f(\theta_t; \xi_t)$ ▷ Stochastic gradient
6: $m_t \leftarrow \rho_t m_{t-1} + (1 - \rho_t) g_t$ ▷ 1^{st}-order momentum
7: $C_t \leftarrow$ Preconditioner construction
8: $\theta_{t+1} \in \underset{\theta \in \Omega}{\operatorname{argmin}} \left\{ \langle m_t, \theta \rangle + \lambda \mathcal{R}(\theta) + \frac{1}{2\alpha_t} (\theta - \theta_t)^\mathsf{T} \big(C_t + \delta I\big)(\theta - \theta_t) \right\}$ ▷ Update rule
9: **end for**

- We analyze the convergence of the general PROXGEN family and identify essential conditions for convergence. We show that in general PROXGEN enjoys the same convergence rate as vanilla SGD, but more importantly that the adaptive methods can have faster convergence in terms of constant than vanilla SGD for sparse data. Our convergence guarantee encompasses several existing approaches as special cases.
- In terms of practice, we demonstrate the superiority of proximal methods over subgradient-based methods with various non-convex regularizers which have not yet been studied in deep learning. Interestingly, our experiments indicate that the benefit of proximal methods over subgradient approaches is more pronounced with non-convex regularizers than with convex regularizers for learning sparse deep models.

Table 1 summarizes the previous studies and our work in terms of stochastic PGD.

2 A Unified Framework of Adaptive Proximal Gradient Methods

In this section, we present PROXGEN, a general family of stochastic proximal gradient methods, and present both existing and novel instances as showcase examples in our family. Algorithm 1 describes the details of PROXGEN. The update rule on line 8 of Algorithm 1 can be written more compactly:

$$\begin{aligned} \theta_{t+1} &\in \underset{\theta \in \Omega}{\operatorname{argmin}} \left\{ \langle m_t, \theta \rangle + \lambda \mathcal{R}(\theta) + \frac{1}{2\alpha_t} (\theta - \theta_t)^\mathsf{T} \Big(C_t + \delta I\Big)(\theta - \theta_t) \right\} \\ &= \operatorname{prox}_{\alpha_t \lambda \mathcal{R}(\cdot)}^{C_t + \delta I} \Big(\theta_t - \alpha_t (C_t + \delta I)^{-1} m_t \Big) \end{aligned} \tag{2}$$

where the proximal operator in (2) is defined as $\operatorname{prox}_h^A(z) = \operatorname{argmin}_x \{h(x) + \frac{1}{2}\|x - z\|_A^2\}$. In PROXGEN, we allow both the loss and the regularizer to be non-convex. Now, we introduce possible examples according to the proper combinations of preconditioners C_t and regularizers $\mathcal{R}(\cdot)$.

Existing Instances of PROXGEN. We briefly recover some known examples in PROXGEN family.

- ADAGRAD [9] is the first key instance of adaptive gradient methods where $C_t = (\sum_{\tau=1}^t g_\tau g_\tau^\mathsf{T})^{1/2}$ and $\mathcal{R}(\theta) = \|\theta\|_1$. Any convex regularizer $\mathcal{R}(\cdot)$ is allowed.
- The proximal Newton methods [29] employ the exact Hessian $C_t = \nabla^2 f(\theta_t)$ and $\mathcal{R}(\theta) = \|\theta\|_1$. In addition, we can approximate the exact Hessian, which yield proximal Newton-*type* methods such as quasi-Newton approximation [30], L-BFGS approximation [31], and adding a multiple of the identity to the Hessian.

Although the above examples enjoy good theoretical properties in convex settings, many of the modern practical optimization problems involve non-convex loss functions such as learning deep models. Moreover, it is known that non-convex regularizers yield better performance (also in terms of theory) than convex penalties in some applications (see [32, 33, 34, 35] and references therein). Considering this motivation and recent advanced optimizers, we arrive at the following new examples.

Novel Instances of PROXGEN. Beyond the well-known methods above, PROXGEN naturally introduces proximal versions of standard SGD techniques developed for solving unregularized problems for deep learning. The following examples are just a few instances that have not been

explored so far, and PROXGEN can cover a broader range of new examples depending on the combinations of preconditioners and regularizers.

- The *proximal version* of ADAM [19] with ℓ_q regularization is a possible example where $C_t = \sqrt{\beta C_{t-1} + (1-\beta) g_t^2}$ with $\beta \in [0, 1)$ and $\mathcal{R}(\theta) = \|\theta\|_q$ for $0 \leq q \leq 1$. We validate empirically the superiority of our novel *proximal version* of ADAM over the usual subgradient-based counterpart in Section 4.
- We can also consider the *proximal* version of KFAC [36]. For an L-layer neural network, KFAC approximates the Fisher information matrix with layer-wise block diagonal structure where l-th diagonal block $C_{t,[l]}$ corresponds to Kronecker-factored approximation with respect to the parameters at l-th layer. The proximal version of KFAC corresponds to $C_{t,[l]} = \mathbb{E}[\boldsymbol{\delta}_l \boldsymbol{\delta}_l^\mathsf{T}] \otimes \mathbb{E}[\boldsymbol{a}_{l-1} \boldsymbol{a}_{l-1}^\mathsf{T}]$ and $\mathcal{R}(\theta) = \|\theta\|_q$ where $\boldsymbol{\delta}_l$ is the gradient with respect to the outputs of l-th layer and $\boldsymbol{a}_{l-1}$ is the activation of $(l-1)$-th layer.

Examples of Proximal Mappings for PROXGEN. We provide update rules for PROXGEN with ℓ_q regularization $(0 \leq q \leq 1)$ and diagonal preconditioners, for which closed-form updates are available. Diagonal preconditioners are used by popular adaptive gradient methods such as ADAM. Note, however, that our framework and convergence analysis are not limited to diagonal preconditioners and apply to general positive preconditioners. Specifically, we consider regularizer $\mathcal{R}(\theta) = \lambda \sum_{j=1}^{p} |\theta_j|^q$ for $\theta \in \mathbb{R}^p$ with diagonal preconditioner matrix C_t. Note that for $C_t = I$ (i.e. vanilla gradient descent), it is known that closed-form solutions exist for proximal mappings for $q \in \{0, \frac{1}{2}, \frac{2}{3}, 1\}$ [37]. We denote the i-th coordinate of the vector θ_t as $\theta_{t,i}$ and the diagonal entry $[C_t]_{ii}$ as $C_{t,i}$

• ℓ_1 **regularization.** The proximal mappings for the case of ℓ_1 regularization with preconditioner can be computed efficiently via soft-thresholding as

$$\widehat{\theta}_{t,i} = \theta_{t,i} - \alpha_t \frac{m_{t,i}}{C_{t,i} + \delta}, \quad \theta_{t+1,i} = \operatorname{sign}\big(\widehat{\theta}_{t,i}\big)\Big(\big|\widehat{\theta}_{t,i}\big| - \frac{\alpha_t \lambda}{C_{t,i} + \delta}\Big) \tag{3}$$

• ℓ_0 **regularization.** In case of ℓ_0 regularization, we can compute the closed-form solutions via hard-thresholding as

$$\widehat{\theta}_{t,i} = \theta_{t,i} - \alpha_t \frac{m_{t,i}}{C_{t,i} + \delta}, \quad \theta_{t+1,i} = \begin{cases} \widehat{\theta}_{t,i}, & |\widehat{\theta}_{t,i}| > \sqrt{\frac{2\alpha_t \lambda}{C_{t,i}+\delta}}, \\ 0, & |\widehat{\theta}_{t,i}| < \sqrt{\frac{2\alpha_t \lambda}{C_{t,i}+\delta}} \\ \{0, \widehat{\theta}_{t,i}\}, & |\widehat{\theta}_{t,i}| = \sqrt{\frac{2\alpha_t \lambda}{C_{t,i}+\delta}} \end{cases} \tag{4}$$

The closed-form proximal mappings for $\ell_{1/2}$ and $\ell_{2/3}$ regularization are provided in the Appendix.

Revised PROXQUANT [1]. The recently proposed PROXQUANT proposes novel regularizations for network quantization. Especially for binary quantization, a W-shaped regularizer is defined as $\mathcal{R}_{\text{bin}}(\theta) = \|\theta - \operatorname{sign}(\theta)\|_1$ where $\operatorname{sign}(\theta)$ is applied on θ in an element-wise manner. Using this regularizer, the main difference between PROXQUANT and our PROXGEN approach is shown in Table 2.

Table 2: PROXQUANT versus *revised* PROXQUANT

PROXQUANT	$\operatorname{prox}_{\alpha_t \lambda \mathcal{R}(\cdot)}\Big(\theta_t - \alpha_t (C_t + \delta I)^{-1} m_t\Big)$
Revised PROXQUANT	$\operatorname{prox}^{C_t + \delta I}_{\alpha_t \lambda \mathcal{R}(\cdot)}\Big(\theta_t - \alpha_t (C_t + \delta I)^{-1} m_t\Big)$

Note that PROXQUANT (top in Table 2) does not consider the effect of preconditioners when computing proximal mappings. Therefore, we revise the proximal update in PROXQUANT by considering preconditioners in proximal mappings with PROXGEN (bottom in Table 2). Moreover, we also propose *generalized regularizers* motivated by ℓ_q regularization for $0 < q < 1$: $\mathcal{R}^q_{\text{bin}}(\theta) = \|\theta - \operatorname{sign}(\theta)\|_q$. In terms of theory, [1] prove the convergence of PROXQUANT only for the *full-batch* gradient with *differentiable* regularizers, which is also guaranteed only for vanilla gradient descent. In contrast, using our *revised* PROXQUANT, we can completely bridge the gap in theory (via Theorem 1 in Section 3, which is stated for *stochastic* optimization), and we provide the *exact* update rule for solving the problem in (1). We also investigate the empirical differences of PROXQUANT and our revised PROXQUANT in Section 4.

3 Convergence Analysis

In this section, we provide convergence guarantees for the PROXGEN family. Our goal is to find an ϵ-stationary point for the problem in (1) where ϵ is the required precision. For notational convenience, we assume that the regularization parameter λ is incorporated into $\mathcal{R}(\theta)$ in (1). To guarantee the convergence under this setting, we should deal with the subdifferentials defined as:

Definition 1 (Fréchet Subdifferential)**.** *Let φ be a real-valued function. The Fréchet subdifferential of φ at $\bar{\theta}$ with $|\varphi(\bar{\theta})| < \infty$ is defined by*

$$\widehat{\partial}\varphi(\bar{\theta}) := \{\theta^* \in \Omega \;\Big|\; \liminf_{\theta \to \bar{\theta}} \frac{\varphi(\theta) - \varphi(\bar{\theta}) - \langle \theta^*, \theta - \bar{\theta} \rangle}{\|\theta - \bar{\theta}\|} \geq 0\}.$$

Definition 2 (Limiting Subdifferential)**.** *Let $\widehat{\partial}\varphi(\bar{\theta})$ be the Fréchet subdifferential in Definition 1. The limiting subdifferential of φ at $\bar{\theta}$ is defined by*

$$\partial\varphi(\bar{\theta}) := \{u \in \mathbb{R}^d : \exists \theta_k \xrightarrow{\varphi} \bar{\theta}, u_k \in \widehat{\partial}\varphi(\theta_k), u_k \to u\}.$$

where $\theta_k \xrightarrow{\varphi} \bar{\theta}$ means $\theta_k \to \bar{\theta}$ with $\varphi(\theta_k) \to \varphi(\bar{\theta})$.

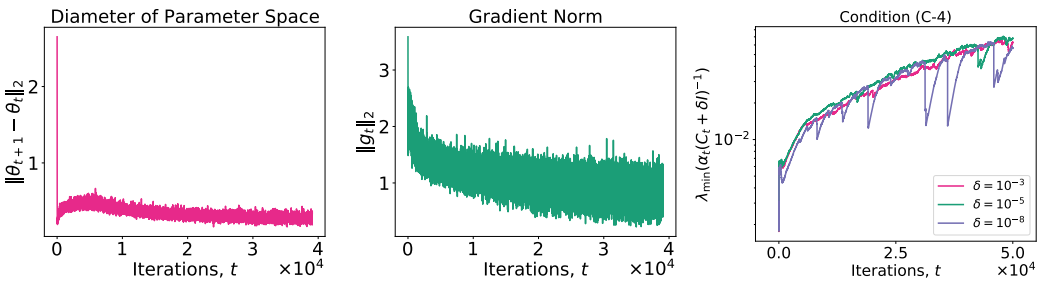
To derive the convergence bound, we make the following mild conditions:

- **(C-1)** (L-smoothness) The loss function f is differentiable, L-smooth, and lower-bounded: $\forall x, y,\ \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ and $f(x^*) > -\infty$ for the optimal solution x^*.
- **(C-2)** (Bounded variance) The stochastic gradient $g_t = \nabla f(\theta_t; \xi)$ is unbiased and has the bounded variance: $\mathbb{E}_\xi\big[\nabla f(\theta_t; \xi)\big] = \nabla f(\theta_t), \quad \mathbb{E}_\xi\big[\|g_t - \nabla f(\theta_t)\|^2\big] \leq \sigma^2$.
- **(C-3)** (i) final step-vector is finite, (ii) the stochastic gradient is bounded, and (iii) the momentum parameter should be exponentially decaying: (i) $\|\theta_{t+1} - \theta_t\| \leq D$, (ii) $\|g_t\| \leq G$, (iii) $\rho_t = \rho_0 \mu^{t-1}$ with $D, G > 0$ and $\rho_0, \mu \in [0, 1)$.
- **(C-4)** (*Sufficiently positive-definite*) The minimum eigenvalue of effective spectrums should be uniformly lower bounded over all time t: $\forall t,\ \lambda_{\min}\big(\alpha_t(C_t + \delta I)^{-1}\big) \geq \gamma > 0$.

(C-1) and (C-2) are very standard in convergence analysis for optimization algorithms designed for deep learning such as ADAM, YOGI, and many others [21, 38, 39, 40, 41]. In addition, (C-3) is extensively studied in previous literature for analysis of general non-convex optimization [19, 20, 42, 40, 41]. Lastly, a similar condition to (C-4) is also considered in [39, 42]. We note that (C-3) and (C-4) are reasonable conditions: It is well-known that the parameter of an overparametrized neural network hardly changes from the initial point during training [43, 44, 45], so one can expect that the diameter D of parameter space and the bound for the size of gradient G have very small values and can be understood as *constants* in rates of the results. To validate this for real cases, we train ResNet-34 on CIFAR-10 dataset. In Figure-1-(a), the difference of parameters $\|\theta_{t+1} - \theta_t\|_2$ and the size of stochastic gradients $\|g_t\|_2$ attain just $1 \sim 3$ while the parameter dimension d of ResNet-34 is about 10^7. Hence, the constants D and G in (C-3) are negligible compared to the problem dimension d in practice. The exponentially decaying momentum parameter assumption $\rho_t = \rho_0 \mu^{t-1}$ could be relaxed to $\rho_t = \rho_0 / t$ sacrificing the logarithmic factor in our analysis. Also, (C-4) is indeed easily satisfied both theoretically and empirically. This condition holds in theory for most of the popular optimization algorithms for deep learning such as ADAGRAD, ADAM, and KFAC (the constant γ is irrelevant to the problem dimension d for each algorithm, and we defer the derivations to Appendix D). In order to investigate whether these conditions could be satisfied in real problems, we revisit the experiments of training ResNet-34. In Figure 1-(b), we can see the minimum eigenvalue of $\alpha_t(C_t + \delta I)^{-1}$ tends to increase, so the condition (C-4) is also satisfied empirically.

Since the loss function f is assumed to be differentiable as in (C-1) and it is known that $\widehat{\partial}\varphi(\theta) \subseteq \partial\varphi(\theta)$, we have, at stationary points, $\mathbf{0} \in \widehat{\partial}F(\theta) = \nabla f(\theta) + \widehat{\partial}\mathcal{R}(\theta)$, so the convergence criterion is slightly different from that of general non-convex optimization. Hence, we use the following convergence criterion $\mathbb{E}[\mathrm{dist}(\mathbf{0}, \widehat{\partial}F(\theta))] \leq \epsilon$ for an ϵ-stationary point where $\mathrm{dist}(x, A)$ denotes the distance between a vector x and a set A. If no regularizer is considered ($\mathcal{R} = 0$), this criterion boils down to the one usually used in non-convex optimization, $\mathbb{E}[\|\nabla f(\theta)\|] \leq \epsilon$.

• **Challenges specific to the analysis of PROXGEN**. The most challenging issue in the analysis of PROXGEN compared to previous studies [14, 21] is that we should handle the momentum m_t and



(a) Bounded domain and bounded gradients (b) Minimum eigenvalue

Figure 1: Empirical results for (a) condition (C-3) and (b) condition (C-4) using ResNet-34.

non-trivial preconditioner C_t. In terms of adaptive gradient methods, [21] guarantees the convergence of a family of adaptive methods (but without proximal mapping) using the changes of effective learning rate ($\Gamma_t := \alpha_t/\sqrt{V_t} - \alpha_{t+1}/\sqrt{V_{t+1}} \geq 0$ where V_t is an adaptation matrix), which is a key quantity in their theory. [21] define a new sequence $\{z_t\}$ involving the quantity Γ_t and exploit the simple closed-form of the quantity $z_{t+1} - z_t$ to derive the convergence with coordinate-wise analysis. However, this proof technique is not available to the *regularized* problems since $z_{t+1} - z_t$ is not amenable anymore to compute in a simple closed-form due to the proximal mapping. On the other hand, our proof *directly* solves the quadratic subproblem w.r.t. θ_t at line 8 in Algorithm 1 to handle a regularizer term. It should also be emphasized that our proof skill can handle arbitrary positive curvatures (hence including more general non-diagonal one) that were not acceptable in [21]. In the context of proximal gradient descent, our proof is totally different from [14] which is only for vanilla SGD. Due to the existence of m_t, it is highly non-trivial to bound the term $\|m_t - \nabla f(\theta_t)\|_2$ without suitable assumptions whereas $\|g_t - \nabla f(\theta_t)\|_2$ in [14] can be easily bounded using (C-2). Also, we need to deal with quadratic approximation term $(\theta - \theta_t)^\top (C_t + \delta I)(\theta - \theta_t)$ in Algorithm 1 which is not problematic in [14] simply because C_t is trivially I. We could successfully bypass those difficulties using mild conditions (C-3) and (C-4), respectively.

We are ready to state our theorem for general convergence.

Theorem 1. *Let θ_a denote an iterate uniformly randomly chosen from $\{\theta_1, \cdots, \theta_T\}$. Under the conditions (C-1), (C-2), (C-3), (C-4) with the initial stepsize $\alpha_0 \leq \frac{\delta}{3L}$ and non-increasing stepsize α_t,* ProxGen*, Algorithm 1, is guaranteed to yield*

$$\mathbb{E}_a[\mathrm{dist}(\mathbf{0}, \widehat{\partial} F(\theta_a))^2] \leq \frac{Q_1 \sigma^2}{T} \sum_{t=0}^{T-1} \frac{1}{b_t} + \frac{Q_2 \Delta}{T} + \frac{Q_3}{T}$$

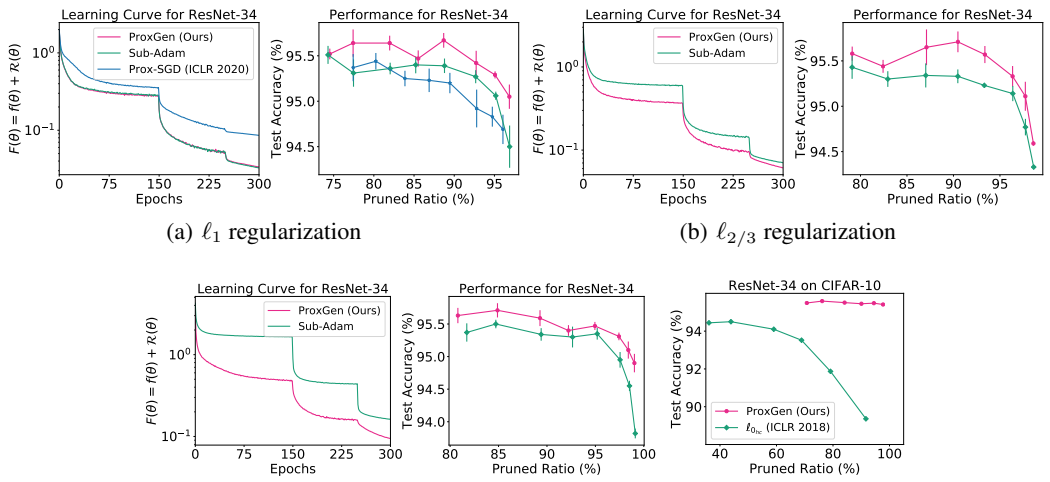
where $\Delta = f(\theta) - f(\theta^*)$ with optimal point θ^*, and b_t is the minibatch size at time t. The constants $\{Q_i\}_{i=1}^3$ on the right-hand side depend on the constants $\{\alpha_0, \delta, L, D, G, \rho_0, \mu, \gamma\}$, but not on T.

Note that the constants $\{Q_i\}_{i=1}^3$ in Theorem 1 are completely independent of the problem dimension d. From Theorem 1, the appropriate minibatch size is important to ensure a good convergence. Various settings for the minibatch size could be employed for convergence guarantee, but considering practical cases, we provide the following important corollary for *constant minibatch*.

Corollary 1 (Constant Mini-batch)**.** *Under the same assumptions as in Theorem 1 with sample size n and constant minibatch size $b_t = b = \Theta(T)$, we have $\mathbb{E}_a\left[\mathrm{dist}(\mathbf{0}, \widehat{\partial} F(\theta_a))^2\right] \leq \mathcal{O}\left(1/T\right)$ and the total complexity is $\mathcal{O}(1/\epsilon^4)$ in order to have $\mathbb{E}_a\left[\mathrm{dist}\left(\mathbf{0}, \widehat{\partial} F(\theta_a)\right)\right] \leq \epsilon$.*

Here we make several remarks on our results and relationship with prior work.

• **On Convergence Results.** Note that our Corollary 1 achieves the optimal complexity $\mathcal{O}(1/\epsilon^4)$ of SGD to find ϵ-stationary points under the standard assumptions (C-1) $\sim$ (C-4). Recent studies [46, 47] show faster rate, but under additional stronger assumptions such as second-order smoothness (i.e., the smoothness of Hessian matrix). Also, we could relax the exponentially decaying momentum $\rho_t = \rho_0 \mu^{t-1}$ in (C-3) to $\rho_t = \rho_0/t$ as mentioned in [48] with the logarithm factor as $Q_3 = \mathcal{O}(\log T)$, which in result still ensures $\widetilde{\mathcal{O}}(1/\epsilon^4)$.



(a) ℓ_1 regularization (b) $\ell_{2/3}$ regularization

(c) $\ell_{1/2}$ regularization (d) ℓ_0 regularization

Figure 2: Comparison for sparse ResNet-34 on CIFAR-10 dataset with step-decay stepsize scheduling.

• **Advantages of using adaptive gradient methods in Theorem 1.** We discuss how the constant γ in (C-4) affects the convergence in terms of theory. According to our proofs, γ depends on the algorithmic details and the constants Q_1, Q_2 and Q_3 in Theorem 1 are proportional to $1/\gamma$. The convergence rate depends on these constants and the benefit of preconditioners can be found here. To view this more clearly, we consider the diagonal matrix adaptation of ADAM [19], i.e. constant stepsize $\alpha_t = \alpha$ and $C_t = \sqrt{(1-\beta)\sum_{\tau=1}^{t} \beta^{t-\tau} g_\tau \odot g_\tau}$, with $\beta \in [0, 1)$ and the total iteration T. In this setting, the $1/\gamma$ can be computed as

$$Q_i \propto \frac{1}{\gamma} = \frac{\sqrt{(1-\beta)\sum_{\tau=1}^{t} \beta^{t-\tau} \|g_\tau\|_2^2} + \delta}{\alpha} \leq \frac{G + \delta}{\alpha}$$

where g_τ is the gradient at time τ and δ is a small constant while the vanilla SGD ($C_t = 0$ and $\delta = 1$) satisfies $1/\gamma = 1/\alpha$. Here, we can clearly see the advantages of adaptive methods (i.e., using preconditioners) since $1/\gamma$ could be dramatically smaller if $\|g_\tau\|_2 \ll 1$ holds roughly with small constant δ, which corresponds to sparse gradients $\|g_\tau\|_2$ (the data features are sparse). This coincides with the convex regret theory for adaptive gradient methods [9, 19, 48], which also holds in our theory with non-convex smooth loss and non-convex regularizers.

• **Implications of condition (C-4) on theory.** Our analysis relies on (C-4), the lower bound for the minimum eigenvalue of $\Gamma_t := \alpha_t (C_t + \delta I)^{-1}$. This means that Theorem 1 guarantees $\mathbb{E}_a[\mathrm{dist}(\mathbf{0}, \widehat{\partial} F(\theta_a)^2] \leq \mathcal{O}(1/\sqrt{T})$ (in case of $b = \Theta(T)$ as in Corollary 1) for *any* change of basis of Γ_t, so in that sense, we provide a worst-case analysis and there is room for more optimistic bounds.

• **On minibatch in Corollary 1.** The conditions $b = \Theta(T)$ is considered as standard in many previous literature [38, 14] and is not stringent. In terms of stochastic optimization, it is natural in practice to choose the batch size b and the number of epochs e in advance. Then, the total number of iterations T satisfies the following relation: $T = e \times \frac{n}{b} = e \times \frac{n}{\Theta(T)}$. In this sense, the total iterations T should be an order of $\mathcal{O}(\sqrt{n})$ in practice. For example, this condition sets a minibatch size of approximately 200 and 1000 for CIFAR-10 and ImageNet dataset respectively, which is practical.

• **Connections to second-order methods.** Our analysis can provide guarantees for *positive* second-order preconditioners as long as (C-4) is satisfied (The empirical Fisher information [36] is one example). Although second-order solvers generally enjoy very fast convergence under strongly convex loss [29, 49], it can be understood that our theory guarantees *at least a sublinear rate for such second-order curvatures* with less stringent conditions.

Table 3: Comparison for binary neural networks. The best performance in mean value is highlighted.

Test Error (%)						
Baselines				PROXGEN (Ours)		
Model	Full Precision (32-bit)	BinaryConnect [8]	PROXQUANT [1]	Revised ProxQuant ℓ_1	Revised ProxQuant $\ell_{2/3}$	Revised ProxQuant $\ell_{1/2}$
ResNet-20	8.06	9.54 ± 0.03	$\mathbf{9.35} \pm 0.13$	9.50 ± 0.12	9.72 ± 0.06	9.78 ± 0.18
ResNet-32	7.25	8.61 ± 0.27	8.53 ± 0.15	8.29 ± 0.07	$\mathbf{8.22} \pm 0.05$	8.43 ± 0.15
ResNet-44	6.96	8.23 ± 0.23	7.95 ± 0.05	$\mathbf{7.68} \pm 0.07$	7.91 ± 0.08	7.90 ± 0.13
ResNet-56	6.54	7.97 ± 0.22	7.70 ± 0.06	$\mathbf{7.52} \pm 0.18$	7.60 ± 0.09	7.61 ± 0.12

4 Experiments

We consider two important tasks for regularized training in deep learning communities: (i) training sparse neural networks and (ii) network quantization. Throughout our experiments, we consider ADAM as a representative of PROXGEN where $m_t = \rho_t m_{t-1} + (1-\rho_t) g_t$ with constant decaying parameter $\rho_t = 0.9$ and $C_t = \sqrt{\beta C_{t-1} + (1-\beta) g_t^2}$ with $\beta = 0.999$ in Algorithm 1. The details on other hyperparameter/experiment settings are provided in the Appendix.

Training Sparse Neural Networks. Motivated by the lottery ticket hypothesis [50], we consider training VGG-16 [51] and ResNet-34 [52] on CIFAR-10 dataset using sparsity encouraging regularizers. Toward this, we consider the following objective function with possibly non-convex ℓ_q regularization: $F(\theta) := \mathbb{E}_{\xi \sim \mathbb{P}}[f(\theta; \xi)] + \lambda \sum_{j=1}^{p} |\theta_j|^q$ where $0 \leq q \leq 1$. We train the network parameters with the closed-form proximal mappings introduced in Section 2. The results on VGG-16 are provided in Appendix.

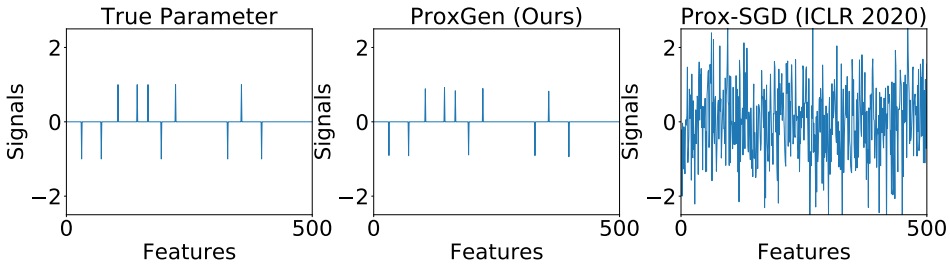
We compare PROXGEN with subgradient methods and also include PROX-SGD [16] as a baseline especially for ℓ_1 regularization since PROX-SGD considers only convex regularizers. In PROX-SGD, the hand-crafted fine-tuned scheduling on α_t and ρ_t is essential for fast convergence and good performance, but in our experiments we use standard settings $\rho_t = 0.9$. We first validate our theory in practice using constant stepsize in order to purely see the effect of proximal approaches (the results on this setting are provided in the Appendix). Then the step-decay learning rate scheduling is employed to consider standard training schemes for the state-of-the-art performance, which also satisfies the non-increasing stepsize condition in our Theorem 1. For ℓ_0 regularization, the problem in (1) cannot be optimized in a subgradient manner, so we compare PROXGEN with another popular baseline, $\ell_{0_{hc}}$ [6] which approximates the ℓ_0-norm via hard-concrete distributions.

Figures 2 illustrates the results for ResNet-34. In terms of convergence, PROXGEN shows faster convergence than PROX-SGD for ℓ_1 case in Figure 2-(a), but there is no difference between PROXGEN and subgradient methods as in Figure 2-(a). However, there are notable differences in convergence for non-convex regularizers $\ell_{1/2}$ and $\ell_{2/3}$, which get bigger as q decreases. We believe this might be because the ℓ_q-norm derivative, $q/|\theta|^{1-q}$, is very large for non-zero tiny θ for $q \in (0, 1)$. Meanwhile, $\partial|\theta|/\partial\theta$ is merely the sign value regardless of size of θ, so the large gradient of $|\theta|^q$ may hinder convergence. The learning curves in Figure 2-(b,c) empirically corroborate this phenomenon.

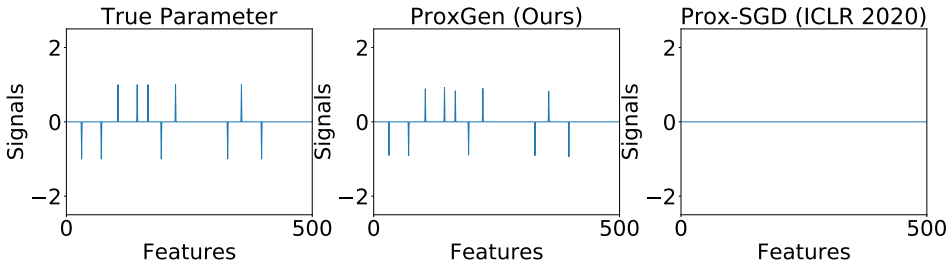
In terms of performance, we can see that PROXGEN consistently achieves better performance than baselines for ResNet-34 with similar or even better sparsity level. Importantly, PROXGEN with ℓ_0 outperforms $\ell_{0_{hc}}$ baseline by a great margin. This might be due to the design of $\ell_{0_{hc}}$, which approximates $\|\theta\|_0 = \sum_{j=1}^{p} \mathbb{I}\{\theta_j \neq 0\}$ with binary mask z_j parameterized by learnable probability π_j for each coordinate. Thus, the number of parameters to be optimized is doubled, which might make optimization harder. In constrast, PROXGEN does not introduce additional parameters.

More results for other famous non-convex regularizers MCP [53] and SCAD [54] are in Appendix.

Training Group-Sparse Neural Networks. In the Appendix, we consider training Statistical Recurrent Units where $\ell_{1,2}$ group-norm penalty is imposed on the input layer weights to detect non-linear Granger Causality [55]. As the proximal mappings for PROXGEN with group sparsity are not available in closed-form, we develop an efficient procedure for computation, whose derivations are also provided in the Appendix.



(a) Random initialization



(b) Zero initialization

Figure 3: Lasso simulations with different initialization schemes.

Training Binary Neural Networks. We consider the network quantization constraining the parameters to some set of discrete values which is a key approach for model compression. We evaluate our revised ProxQuant in Table 2 with extended regularization $\mathcal{R}^q_{\text{bin}}$ in Section 2. We consider the following objective function with quantization-specific regularizers: $F(\theta) := \mathbb{E}_{\xi\sim\mathbb{P}}[f(\theta;\xi)] + \lambda \sum_{j=1}^{p} |\theta_j - \text{sign}(\theta_j)|^q$ where $0 \le q \le 1$. For comparisons, we quantize ResNet weight parameters (except bias and activations) on CIFAR-10 and ImageNet dataset.

Table 3 presents the results. For all q values, revised ProxQuant consistently outperforms the baselines except for ResNet-20, which implies ProxGen may work better for larger networks. As such, our generalized regularizers $\mathcal{R}^q_{\text{bin}}$ contribute to one of the state-of-the-art optimization-based methods in network quantization. Notably, revised ProxQuant ℓ_1 greatly outperforms ProxQuant baseline while these two approaches differ only in update rules (see Table 2). Hence, we can conclude that revised ProxQuant based on ProxGen provides an *exact* proximal update and also yields more generalizable solutions. In our experience, revised ProxQuant ℓ_0 shows little degradation in performance, so we do not include this result. However, revised ProxQuant ℓ_0 shows superiority to baselines for language modeling, whose preliminary results are in Appendix.

Table 4: Comparison for binary neural networks for ImageNet. † means the first and last layer not quantized.

	ResNet-18	
	Top-1 Error (%)	Top-5 Error (%)
Full precision	30.46	10.81
BWN [56]	39.20	17.00
LR-Net† [57]	40.10	17.70
ELQ [58]	35.28	13.96
ProxQuant [1]	36.24	14.23
Revised ProxQuant ℓ_1 (Ours)	**34.85**	**12.38**

Table 4 illustrates Top-1/Top-5 error (%) for training ResNet on ImageNet with binary quantization. The most important thing is that our revised ProxQuant shows great improvements in performance over the original ProxQuant. Furthermore, ProxGen shows superior performance to various baselines for weight quantization.

5 A Closer Look into Prox-SGD [16] vs. ProxGen

Prox-SGD [16] is the approach closest to our ProxGen method. However, Prox-SGD is *not an exact* proximal approach and is significantly different from ProxGen. ProxGen's update rule

involves directly solving the quadratic subproblem (2). In contrast, PROX-SGD's update rule consists of two stages: (i) solving the quadratic subproblem *without* learning rate (5), then (ii) updating the parameters with the computed direction (i.e. $\widehat{\theta}_t - \theta_t$) by the learning rate α_t (6).

$$\widehat{\theta}_t = \underbrace{\text{prox}_{\lambda\mathcal{R}(\cdot)}^{C_t+\delta I}\Big(\theta_t - (C_t+\delta I)^{-1} m_t\Big)}_{\text{no learning rate}}, \tag{5}$$

$$\theta_{t+1} = \theta_t + \alpha_t(\widehat{\theta}_t - \theta_t) \tag{6}$$

To clearly see the differences between both approaches, we conduct two studies.

Study 1: Lasso Support Recovery. For this task, the two-stage update scheme of PROX-SGD might have some potential issues. For example, for ℓ_1-regularized problems, the updated parameter θ_{t+1} (6) *might not achieve exact zero* (while $\widehat{\theta}_t$ can) whereas θ_{t+1} for PROXGEN (2) can attain exact zero value according to the update rule (3) in Section 2. Another potential caveat is that PROX-SGD might *overestimate* the sparsity level. In view of the above, we run Lasso simulations with different two initialization schemes: (i) random initialization and (ii) zero initialization. For random initialization, it can be seen in Figure 3-(a) that PROX-SGD could not achieve exact zero value, which corroborates our first observation. More interestingly, for zero initialization, we can see in Figure 3-(b) that the estimates using PROX-SGD are exactly zeros for all coordinates, which supports our second observation. This might be because $\widehat{\theta}_t$ (5) is always zero since the quadratic subproblem does not consider the learning rate, which might overestimate the sparsity level. Hence, the subsequent iterate θ_{t+1} would be always zero since we initialize the parameters with zeros, but PROXGEN recovers the correct support in both cases.

Study 2: DenseNet-201 on CIFAR-100 Dataset. To validate the superiority of PROXGEN upon PROX-SGD, we revisit the largest experiments in [16]. We train DenseNet-201 architecture on CIFAR-100 dataset with ℓ_1 regularization since PROX-SGD only consider convex regularizers. For both methods, we use the same hyperparameter settings for fair comparison. Figure 4 illustrates the training learning curves, and it can be seen that our PROXGEN achieves faster convergence as well as lower objective values. For our experience, the learning curves show the similar dynamics for different λ values.

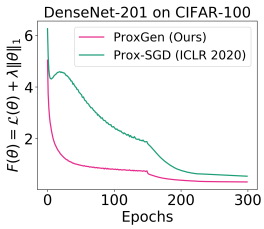


Figure 4: Learning curve.

Comparison of Theoretical Contributions. [16] guarantees the convergence of PROX-SGD, but *not how fast* it converges. Moreover, this is proved *without* considering preconditioners. In contrast, our analysis for the PROXGEN framework appropriately incorporates the first-order momentum and arbitrary positive preconditioner with detailed *non-asymptotic* convergence.

6 Conclusion

In this work, we proposed PROXGEN, the first general family of stochastic proximal gradient methods. Within our framework, we presented novel examples of proximal versions of standard SGD approaches, including a proximal version of ADAM. We analyzed the convergence of the whole PROXGEN family and showed that PROXGEN can encompass the results of several previous studies. We also demonstrated that PROXGEN empirically outperforms subgradient-based methods for popular deep learning problems. As future work, we plan to further study efficient procedures to compute the proximal mappings for structured regularizers such as ℓ_1/ℓ_q-norms with preconditioners.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grants (2018R1A5A1059921, 2019R1C1C1009192) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence, No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)) funded by the Korea government (MSIT).

References

[1] Yu Bai, Yu-Xiang Wang, and Edo Liberty. Proxquant: Quantized neural networks via proximal operators. In *International Conference on Learning Representations*, 2019.

[2] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[3] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[4] A. N. Tychonoff. On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39(5):195–198, 1943.

[5] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In *Advances in neural information processing systems*, pages 2074–2082, 2016.

[6] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018.

[7] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. Quantization networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7308–7316, 2019.

[8] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pages 3123–3131, 2015.

[9] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Journal of Machine Learning Research (JMLR)*, 2011.

[10] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[11] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

[12] Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *arXiv preprint arXiv:1902.05679*, 2019.

[13] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[14] Yi Xu, Rong Jin, and Tianbao Yang. Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. In *Advances in Neural Information Processing Systems*, pages 2626–2636, 2019.

[15] Yi Xu, Qi Qi, Qihang Lin, Rong Jin, and Tianbao Yang. Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence. In *International conference on machine learning*, 2019.

[16] Yang Yang, Yaxiong Yuan, Avraam Chatzimichailidis, Ruud JG van Sloun, Lei Lei, and Symeon Chatzinotas. Proxsgd: Training structured neural networks under regularization and constraints. In *International Conference on Learning Representations*, 2020.

[17] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[18] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation (ICLR)*, 2015.

[20] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[21] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[22] Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[23] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[25] Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pages 1145–1153, 2016.

[26] Zeyuan Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 89–97. JMLR. org, 2017.

[27] Tianyi Chen, Tianyu Ding, Bo Ji, Guanyi Wang, Yixin Shi, Sheng Yi, Xiao Tu, and Zhihui Zhu. Orthant based proximal stochastic gradient method for ℓ_1-regularized optimization. *arXiv preprint arXiv:2004.03639*, 2020.

[28] Emilie Chouzenoux, Jean-Christophe Pesquet, and Audrey Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132, 2014.

[29] Jason D Lee, Yuekai Sun, and Michael Saunders. Proximal newton-type methods for convex optimization. In *Advances in Neural Information Processing Systems*, pages 827–835, 2012.

[30] Stephen Becker, Jalal Fadili, and Peter Ochs. On quasi-newton forward-backward splitting: Proximal calculus and convergence. *SIAM Journal on Optimization*, 29(4):2445–2481, 2019.

[31] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[32] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[33] Cheolwoo Park and Young Joo Yoon. Bridge regression: adaptivity and group selection. *Journal of Statistical Planning and Inference*, 141(11):3506–3519, 2011.

[34] Eunho Yang and Aurélie C Lozano. Sparse+ group-sparse dirty models: Statistical guarantees without unreasonable conditions and a case for non-convexity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3911–3920. JMLR. org, 2017.

[35] Jihun Yun, Peng Zheng, Eunho Yang, Aurelie Lozano, and Aleksandr Aravkin. Trimming the ℓ_1 regularizer: Statistical analysis, optimization, and applications to deep learning. In *International Conference on Machine Learning*, pages 7242–7251, 2019.

[36] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417, 2015.

[37] Wenfei Cao, Jian Sun, and Zongben Xu. Fast image deconvolution using closed-form thresholding formulas of lq (q= 12, 23) regularization. *Journal of visual communication and image representation*, 24(1):31–41, 2013.

[38] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in neural information processing systems*, pages 9793–9803, 2018.

[39] Jihun Yun, Aurelie C. Lozano, and Eunho Yang. Stochastic gradient methods with block diagonal matrix adaptation. *arXiv preprint arXiv:1905.10757*, 2019.

[40] Zaiyi Chen, Zhuoning Yuan, Jinfeng Yi, Bowen Zhou, Enhong Chen, and Tianbao Yang. Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *International Conference on Learning Representations*, 2019.

[41] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pages 6677–6686. PMLR, 2019.

[42] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019.

[43] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[44] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[45] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[46] Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[47] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[48] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[49] Guodong Zhang, James Martens, and Roger B Grosse. Fast convergence of natural gradient descent for over-parameterized neural networks. In *Advances in Neural Information Processing Systems*, pages 8080–8091, 2019.

[50] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[53] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[54] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[55] Saurabh Khanna and Vincent Y. F. Tan. Economy statistical recurrent units for inferring nonlinear granger causality. In *International Conference on Learning Representations*, 2020.

[56] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016.

[57] Oran Shayer, Dan Levi, and Ethan Fetaya. Learning discrete weights using the local reparameterization trick. In *International Conference on Learning Representations*, 2018.

[58] Aojun Zhou, Anbang Yao, Kuan Wang, and Yurong Chen. Explicit loss-error-aware quantization for low-bit deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9426–9435, 2018.

[59] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

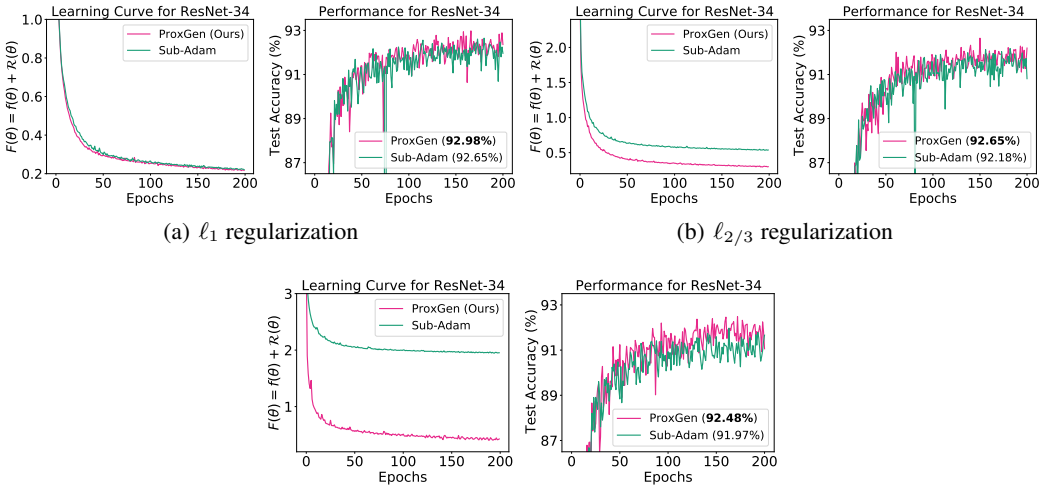
[60] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[61] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Supplementary Materials

A Sparse Neural Networks with ℓ_q Regularization with Constant Stepsize

We include the experimental results on sparse neural networks using ResNet-34 and constant stepsize in Figure 5. As seen in Figure 5, the proximal methods in this regime also shows superior performance than the subgradient baselines.



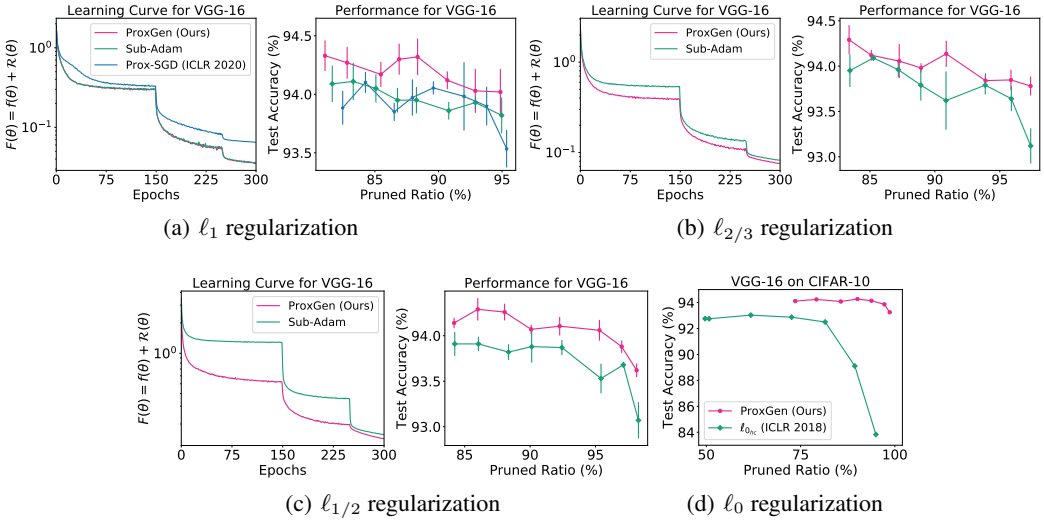
(a) ℓ_1 regularization (b) $\ell_{2/3}$ regularization

(c) $\ell_{1/2}$ regularization

Figure 5: Comparison for sparse ResNet-34 on CIFAR-10 dataset using constant stepsize.

B Sparse Neural Networks with ℓ_q Regularization for VGG-16

We include the experimental results on sparse neural networks using VGG-16 architecture in Figure 6.



(a) ℓ_1 regularization (b) $\ell_{2/3}$ regularization

(c) $\ell_{1/2}$ regularization (d) ℓ_0 regularization

Figure 6: Comparison for sparse VGG-16 on CIFAR-10 dataset.

Table 5: Configuration and Hyperparameters for the SRU experiments.

Parameters	Dataset				
	Lorenz F=10	Lorenz F=40	VAR	Dream-3	NetSim
Learning rate	0.005	0.01	0.04	0.005	0.001
Batch size	125	125	125	21	5
# Training epochs	2000	2000	2000	1000	2000
# units per layer	10				
$\mathcal{A}$	$\{0.0, 0.01, 0.1, 0.99\}$				
Group-sparse reg. param. for the input layer	$[0.01, 10]$				

C Additional Experiments: Sparse Neural Networks with MCP and SCAD Non-convex Regularizers

We provide the additional experiments for sparse neural networks with MCP [53] and SCAD [54] non-convex regularizers. Figure 7 and 8 illustrate the results for VGG-16 and ResNet-34 respectively. As shown in Section 4 and these figures, PROXGEN is very effective for solving the non-convex regularized problems.

D Experiments on Group-Sparse Neural Networks

In this section, we consider estimating group-sparse Neural networks for the task of detecting non-linear Granger Causality in multivariate time series data. Granger causality [59] is a widely used approach for time series structure discovery. Several approaches have been proposed recently that employ structured multilayer perceptrons (MLPs) or recurrent neural networks (RNNs) [60, 55]. Following [55] (Section 3), given n time series, for each time series x_j, we consider training a Statistical Recurrent Unit network to predict the future value x_j based on the past value of all the n time series. The $\ell_{1,2}$ group-norm penalty is imposed on the input layer parameters where group G_i is formed by the input layer parameters corresponding to input time series x_i. Then, time series x_i is detected as Granger-causing time series x_j if the input layer parameters in group G_i for the SRU network predicting x_j are non-zero.

Comparison methods. For optimization of the group-sparsity regularized SRUs, [55] employs ADAM where each ADAM step is followed by a proximal step via group soft-thresholding of the input layer parameters. This proximal step uses ADAM's initial learning rate and disregards its coordinate-wise adaptivity, hence the optimization method of [55] is not a "pure" proximal stochastic gradient descent approach. In this section, we compare the algorithm of [55](using the implementation from https://github.com/sakhanna/SRU_for_GCI) with PROXGEN. The proximal updates for PROXGEN with the $\ell_{1,2}$ group-norm penalty are not available in closed-form and we use the procedure described in Section F to compute them.

Evaluation metrics. We follow the same experimental setup as in [55] and evaluate the methods in terms of their accuracy in detecting causal links among time series. Specifically we report the AUROC (Area Under the Receiver Operating Characteristic curve), where the ROC curve illustrates the trade off between the true-positive rate (TPR) and the false-positive rate (FPR) achieved by the methods towards the detection of pairwise Granger causal relationships.

Datasets. We use the same datasets as in [55] : *Lorenz* (F=10/40) where $T = \{250, 500\}$ measurements for 10 time series variables are generated according to the Lorenz-96 model and F denotes the magnitude of the external forcing in the model; *VAR* simulations for a 3rd order VAR model with 10 components and $T = \{500, 1000\}$ measurements ; *NetSim* where we use $T = 200$ time-ordered signals that are simulated for 15 brain regions in 5 human subjects labelled $2 - 6$; and *Dream-3 (E.coli-1)* where gene expression levels for 100 genes are measured for *E.coli* over 966 time points. The datasets are all available from https://github.com/sakhanna/SRU_for_GCI.

Hyperparameters and configuration. We use the same hyperparameters and SRU configuration as [55] (Table 10). The only difference is that for PROGEN we set the range of the penalty parameter for the group penalty to $[0.001, 10]$. The setups are summarized in Table 5.

Results. The AUROC for the various datasets are presented in Table 6. As can be seen from the table, PROXGEN achieves higher AUROC in most cases. We believe that the improved performance of PROXGEN is

Table 6: Comparison for group-sparsity regularized SRUs. AUROC (the higher the better) for detecting pairwise causal links from multivariate time series data.

	[55]	PROXGEN
Lorenz ($F = 10, T = 250$)	0.83 ± 0.03	$\mathbf{0.94 \pm 0.02}$
Lorenz ($F = 10, T = 500$)	0.90 ± 0.02	$\mathbf{0.98 \pm 0.05}$
Lorenz ($F = 40, T = 250$)	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$
Lorenz ($F = 40, T = 500$)	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$
VAR (T=500)	0.82 ± 0.06	$\mathbf{0.88 \pm 0.04}$
VAR (T=1000)	0.91 ± 0.04	$\mathbf{0.93 \pm 0.05}$
NetSim	$\mathbf{0.79 \pm 0.03}$	0.78 ± 0.02
Dream-3 (E.coli-1)	0.657	**0.660**

directly attributable to the incorporation of the preconditionners in the proximal mappings. As future work, we plan to experiment with other architectures beyond SRUs, such as the Economy SRUs proposed in [55].

E Details on Experimental Settings

Sparse Neural Networks. To reflect the most practical training settings, we first tune the weight-decay parameter ζ without ℓ_q regularizers. For weight-decay coefficients, we consider the candidates $\zeta \in \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ for ζ and the best ζ value is 0.2 for both networks VGG-16 and ResNet-34 in our experience. After tuning weight-decay coefficient ζ, we consider both decoupled weight decay [61] and ℓ_q regularization whose detail update rule is described in Algorithm 2. For all comparison methods except $\ell_{0_{hc}}$, the recommended stepsize $\alpha_t = 0.001$ is employed, but we tune this stepsize for $\ell_{0_{hc}}$ baseline. We consider a broad range of regularization parameters for all methods: $\lambda \in \{0.001, 0.002, 0.005, 0.01, 0.02, \cdots, 1.0, 2.0, 5.0\}$. With these hyperparameter settings, we consider the total 300 epochs and divide the learning rate at 150-th and 250-th epoch by 10.

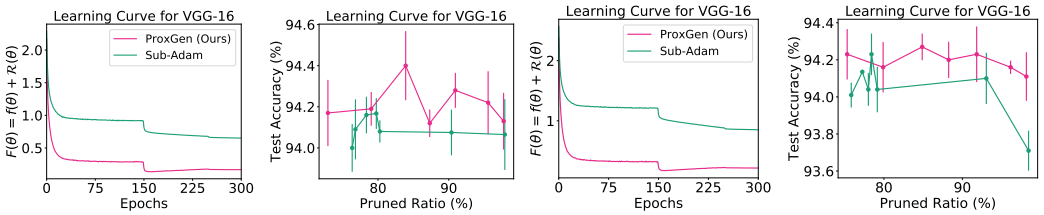
Binary Neural Networks. In this experiment, we follow the same experimental settings in baseline PROXQUANT [1]. We first pre-train ResNet-{20, 32, 44, 56} with full-precision and initialize the network parameters with these pre-trained weights. Then, we consider the total 300 epochs and hard-quantize the networks at 200-th epoch (i.e. quantizing the weight parameters to $+1$ or -1). We employ the homotopy method introduced in [1]: annealing the regularization paramter λ as $\lambda_{\text{epoch}} = \lambda \times$ epoch. For initial value of λ, we use $\lambda = 10^{-8}$ or $\lambda = 5 \cdot 10^{-8}$ for all ResNet architecture. We use the constant stepsize $\alpha_t = 0.01$ as recommended in [1].

Lasso Support Recovery. We generate simple Lasso simulations with problem dimension $p = 500$ and $n = 100$ data samples. The number of non-zero entries in true parameter vector $\theta^* \in \mathbb{R}^p$ is set to 10. The design matrix $X \in \mathbb{R}^{n \times p}$ is generated from standard Gaussian distribution $\mathcal{N}(0, 1)$ and we randomly assign $+1$ or -1 for the non-zero value in true parameter at random 10 coordinates. The response variable $y \in \mathbb{R}^n$ is generated with small noise by $y = X\theta^* + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.05^2)$. For both PROXGEN and PROX-SGD, we employ ADAM for preconditioner matrix C_t construction.

Here, we introduce preliminary results of revised PROXQUANT ℓ_0 on language modeling. For this experiment, we train one hidden layer LSTM with embedding dimension 300 and 300 hidden units according to [1]. First, we pre-train the full-precision LSTM and initialize the network with pre-trained weights. We consider the total 80 epochs and divide the learning rate by 1.2 if the validation loss does not decrease. Table 7 shows the preliminary results and revised PROXQUANT ℓ_0 is superior to the PROXQUANT baseline in this task.

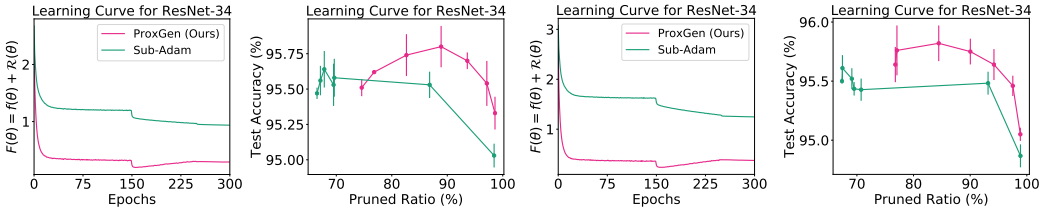
Table 7: Preliminary results on revised PROXQUANT ℓ_0 for LSTM models.

Algorithm	Test Perplexity
Full-precision (32-bit)	88.5
BinaryConnect [8]	372.2
PROXQUANT [1]	288.5
revised PROXQUANT ℓ_0 (Ours)	**223.4**



(a) MCP regularization (b) SCAD regularization

Figure 7: Comparison for sparse VGG-16 on CIFAR-10 dataset with other non-convex regularizers.



(a) MCP regularization (b) SCAD regularization

Figure 8: Comparison for sparse ResNet-34 on CIFAR-10 dataset with other non-convex regularizers.

F Derivations for Proximal Mappings

We first derive the concrete update rule for ℓ_q regularization with *diagonal* preconditioners as introduced in Section 2. Next, we provide closed-form proximal mappings for MCP and SCAD regularization. Finally, we consider group $\ell_{1,2}$ regularization.

$\ell_{1/2}$ regularization. First, we review the closed-form proximal mappings for $\ell_{1/2}$ regularization of vanilla SGD. First, we consider the following one-dimensional program:

$$\widehat{x} = \operatorname*{argmin}_x \{(x-z)^2 + \lambda |x|^{1/2}\} \tag{7}$$

For the program (7), it is known that the closed-form solution exists [37] as

$$\widehat{x} = \begin{cases} \frac{2}{3}|z|\Big(1+\cos\big(\frac{2}{3}\pi - \frac{2}{3}\varphi_\lambda(z)\big)\Big) & \text{if } z > p(\lambda) \\ 0 & \text{if } |z| \le p(\lambda) \\ -\frac{2}{3}|z|\Big(1+\cos\big(\frac{2}{3}\pi - \frac{2}{3}\varphi_\lambda(z)\big)\Big) & \text{if } z < -p(\lambda) \end{cases} \tag{8}$$

where $\varphi_\lambda(z) = \arccos\left(\frac{\lambda}{8}\left(\frac{|z|}{3}\right)^{-3/2}\right)$ and $p(\lambda) = \frac{\sqrt[3]{54}}{4}(\lambda)^{2/3}$. Based on this closed-form solution, we derive ProxGen for $\ell_{1/2}$ regularization with diagonal preconditioners. By (2), we have

$$\widehat{\theta}_t = \theta_t - \alpha_t (C_t + \delta I)^{-1} m_t \tag{9}$$

$$\theta_{t+1} \in \operatorname{prox}^{C_t+\delta I}_{\alpha_t \lambda \mathcal{R}(\cdot)}(\widehat{\theta}_t) \tag{10}$$

$$= \operatorname*{argmin}_\theta \left\{ \frac{1}{2}\|\theta - \widehat{\theta}_t\|^2_{C_t+\delta I} + \lambda \sum_{j=1}^{p} |\theta_j|^{1/2} \right\} \tag{11}$$

Since the program (11) is coordinate-wise decomposable (since the preconditioner matrix C_t is diagonal), we can split (11) into

$$\theta_{t+1,i} = \operatorname*{argmin}_{\theta_i} \left\{ \frac{1}{2}(C_{t,i}+\delta)(\theta_i - \widehat{\theta}_{t,i})^2 + \alpha_t \lambda |\theta_i|^{1/2} \right\}$$

$$= \operatorname*{argmin}_{\theta_i} \left\{ (\theta_i - \widehat{\theta}_{t,i})^2 + \frac{2\alpha_t \lambda}{C_{t,i}+\delta} |\theta_i|^{1/2} \right\}$$

for the i-th coordinate. From (7), we can derive

$$\theta_{t+1,i} = \begin{cases} \frac{2}{3}|\widehat{\theta}_{t,i}|\Big(1+\cos\big(\frac{2}{3}\pi - \frac{2}{3}\varphi_\lambda(\widehat{\theta}_{t,i})\big)\Big) & \text{if } \widehat{\theta}_{t,i} > p(\lambda) \\ 0 & \text{if } |\widehat{\theta}_{t,i}| \le p(\lambda) \\ -\frac{2}{3}|\widehat{\theta}_{t,i}|\Big(1+\cos\big(\frac{2}{3}\pi - \frac{2}{3}\varphi_\lambda(\widehat{\theta}_{t,i})\big)\Big) & \text{if } \widehat{\theta}_{t,i} < -p(\lambda) \end{cases}$$

where

$$\varphi_\lambda(\widehat{\theta}_{t,i}) = \arccos\Big(\frac{\alpha_t\lambda}{4(C_{t,i}+\delta)}\big(\frac{|\widehat{\theta}_{t,i}|}{3}\big)^{-3/2}\Big), \quad p(\lambda) = \frac{\sqrt[3]{54}}{4}\Big(\frac{2\alpha_t\lambda}{C_{t,i}+\delta}\Big)^{2/3}.$$

$\ell_{2/3}$ regularization. Now, we provide the closed-form solutions for proximal $\ell_{2/3}$ mappings with diagonal preconditioners. Similar to $\ell_{1/2}$ regularization, we start from the closed-form solutions of the following program:

$$\widehat{x} = \operatorname*{argmin}_x\{(x-z)^2 + \lambda|x|^{2/3}\} \tag{12}$$

The closed-form solution for the program (12) is known to be

$$\widehat{x} = \begin{cases} \left(\frac{|A|+\sqrt{\frac{2|z|}{|A|}-|A|^2}}{2}\right)^3 & \text{if } z > \frac{2}{3}\sqrt[4]{3\lambda^3} \\ 0 & \text{if } |z| \le \frac{2}{3}\sqrt[4]{3\lambda^3} \\ -\left(\frac{|A|+\sqrt{\frac{2|z|}{|A|}-|A|^2}}{2}\right)^3 & \text{if } z < -\frac{2}{3}\sqrt[4]{3\lambda^3} \end{cases} \tag{13}$$

where

$$|A| = \frac{2}{\sqrt{3}}\lambda^{1/4}\Big(\cosh\big(\frac{\phi}{3}\big)\Big)^{1/2}, \quad \phi = \operatorname{arccosh}\Big(\frac{27z^2}{16}\lambda^{-3/2}\Big) \tag{14}$$

Based on this formulation, we derive the closed-form proximal mappings with diagonal preconditioner C_t. By (2), we have

$$\widehat{\theta}_t = \theta_t - \alpha_t(C_t+\delta I)^{-1}m_t \tag{15}$$

$$\theta_{t+1} \in \operatorname{prox}^{C_t+\delta I}_{\alpha_t\lambda\mathcal{R}(\cdot)}(\widehat{\theta}_t) \tag{16}$$

$$= \operatorname*{argmin}_\theta\Big\{\frac{1}{2}\|\theta-\widehat{\theta}_t\|^2_{C_t+\delta I} + \lambda\sum_{j=1}^p|\theta_j|^{2/3}\Big\} \tag{17}$$

As in $\ell_{1/2}$ case, the program (17) is coordinate-wise separable, so it suffices to solve the sub-problems for each coordinate as

$$\begin{aligned}\theta_{t+1,i} &= \operatorname*{argmin}_{\theta_i}\Big\{\frac{1}{2}(C_{t,i}+\delta)(\theta_i-\widehat{\theta}_i)^2 + \alpha_t\lambda|\theta_i|^{2/3}\Big\} \\ &= \operatorname*{argmin}_{\theta_i}\Big\{(\theta_i-\widehat{\theta}_{t,i})^2 + \frac{2\alpha_t\lambda}{C_{t,i}+\delta}|\theta_i|^{2/3}\Big\}\end{aligned}$$

From (12), we can derive

$$\theta_{t+1,i} = \begin{cases} \left(\frac{|A|+\sqrt{\frac{2|\widehat{\theta}_{t,i}|}{|A|}-|A|^2}}{2}\right)^3 & \text{if } \widehat{\theta}_{t,i} > \frac{2}{3}\sqrt[4]{3\lambda^3} \\ 0 & \text{if } |\widehat{\theta}_{t,i}| \le \frac{2}{3}\sqrt[4]{3\lambda^3} \\ -\left(\frac{|A|+\sqrt{\frac{2|\widehat{\theta}_{t,i}|}{|A|}-|A|^2}}{2}\right)^3 & \text{if } \widehat{\theta}_{t,i} < -\frac{2}{3}\sqrt[4]{3\lambda^3} \end{cases}$$

where

$$|A| = \frac{2}{\sqrt{3}}\Big(\frac{2\alpha_t\lambda}{C_{t,i}+\delta}\Big)^{1/4}\Big(\cosh\big(\frac{\phi}{3}\big)\Big)^{1/2}, \quad \phi = \operatorname{arccosh}\Big(\frac{27\widehat{\theta}^2_{t,i}}{16}\Big(\frac{2\alpha_t\lambda}{C_{t,i}+\delta}\Big)^{-3/2}\Big)$$

In addition to ℓ_q regularization, we provide the closed-form proximal mappings for MCP and SCAD regularizers with non-trivial preconditioners.

MCP regularization. Before introducing the closed-form of proximal m mappings for MCP regularized problems with diagonal preconditioners, we first review the MCP regularizer. The MCP regularizer is defined as

$$\rho_\lambda(x;b) = \begin{cases} \lambda|x| - \frac{x^2}{2b} & \text{if } |x| \le b\lambda \\ \frac{b\lambda^2}{2} & \text{if } |x| > b\lambda \end{cases} \tag{18}$$

where $b > 0$ is called the MCP parameter and λ is a regularization parameter. Our goal is to derive the proximal mapping of this regularizer with diagonal preconditioner.

Now, we start from the closed-form solutions of the following program:

$$\widehat{x} = \operatorname*{argmin}_x \{\frac{1}{2}(x-z)^2 + \rho_\lambda(x;b)\} \tag{19}$$

For this program, the closed-form solution is known as

$$\widehat{x} = \operatorname{sign}(z)\min\left\{\frac{b\max\{|z|-\lambda,0\}}{b-1}, |z|\right\} \tag{20}$$

Based on this closed-form solution, we derive the closed-form proximal mappings with diagonal preconditioner C_t. By (2), we have

$$\widehat{\theta}_t = \theta_t - \alpha_t(C_t+\delta I)^{-1}m_t \tag{21}$$

$$\theta_{t+1} \in \operatorname{prox}^{C_t+\delta I}_{\alpha_t\rho_\lambda(\cdot;b)}(\widehat{\theta}_t) \tag{22}$$

$$= \operatorname*{argmin}_\theta \left\{\frac{1}{2}\|\theta-\widehat{\theta}_t\|^2_{C_t+\delta I} + \alpha_t\rho_\lambda(\theta;b)\right\} \tag{23}$$

Since this program is also coordinate-wise separable, we could have for each coordinate

$$\theta_{t+1,i} = \operatorname{sign}(\widehat{\theta}_{t,i})\min\left\{\frac{b\max\{|\widehat{\theta}_{t,i}| - \frac{\alpha_t\lambda}{C_{t,i}+\delta},0\}}{b-1}, |\widehat{\theta}_{t,i}|\right\} \tag{24}$$

SCAD regularization. We first introduce SCAD regularizer defined as :

$$\rho_\lambda(x;a) = \begin{cases} \lambda|x| & \text{if } |x| \le \lambda \\ \frac{-\lambda^2-2a\lambda|x|+x^2}{2(a-1)} & \text{if } \lambda < |x| \le a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |x| > a\lambda \end{cases} \tag{25}$$

where $a > 2$ is called the SCAD parameter and λ is a regularization parameter. As in MCP regularizer, we start from the following program

$$\widehat{x} = \operatorname*{argmin}_x \left\{\frac{1}{2}\|x-z\|^2 + \rho_\lambda(x;a)\right\}$$

The closed-form solution for this program is known as

$$\widehat{x} = \begin{cases} \operatorname{sign}(z)\max\{|z|-\lambda,0\} & \text{if } |z| \le 2\lambda \\ \frac{(a-1)z-\operatorname{sign}(z)a\lambda}{a-2} & \text{if } 2\lambda < |z| \le a\lambda \\ z & \text{if } |z| > a\lambda \end{cases} \tag{26}$$

Based on this formulation, we could derive the closed-form solution for ProxGen with diagonal preconditioner. By (2), we have

$$\widehat{\theta}_t = \theta_t - \alpha_t(C_t+\delta I)^{-1}m_t \tag{27}$$

$$\theta_{t+1} \in \operatorname{prox}^{C_t+\delta I}_{\alpha_t\rho_\lambda(\cdot;a)}(\widehat{\theta}_t) \tag{28}$$

$$= \operatorname*{argmin}_\theta \left\{\frac{1}{2}\|\theta-\widehat{\theta}_t\|^2_{C_t+\delta I} + \alpha_t\rho_\lambda(\theta;a)\right\} \tag{29}$$

Since the program is coordinate-wise decomposable, we have for each coordinate

$$\theta_{t+1,i} = \begin{cases} \operatorname{sign}(\widehat{\theta}_{t,i})\max\{|\widehat{\theta}_{t,i} - \widehat{\lambda}_i|,0\} & \text{if } |\widehat{\theta}_{t,i}| \le 2\widehat{\lambda}_i \\ \frac{(a-1)\widehat{\theta}_{t,i}-\operatorname{sign}(\widehat{)}\theta_{t,i}a\widehat{\lambda}_i}{a-2} & \text{if } 2\widehat{\lambda}_i < |\widehat{\theta}_{t,i}| \le a\widehat{\lambda}_i \\ \widehat{\theta}_{t,i} & \text{if } |\widehat{\theta}_{t,i}| > a\widehat{\lambda}_i \end{cases} \tag{30}$$

where $\widehat{\lambda}_i = \frac{\alpha_t\lambda}{C_{t,i}+\delta}$.

Although the derivations look little complicated for both cases, we emphasize that both two closed-form solutions can be efficiently implemented in a GPU-friendly manner.

Group $\ell_{1,2}$ regularization. When there is no preconditioning, the proximal mapping for the $\ell_{1,2}$ penalty can be computed in closed-form via group soft-thresholding. In the presence of preconditioners, the proximal mapping for the $\ell_{1,2}$ group penalty is no longer available in closed-form, but can be computed easily as follows.

Let θ denote the network parameters that are being regularized via group penalty, let $\{G_1, \dots, G_K\}$ denote their partition into K groups, and let $\theta_{(k)}$ denote the subset of the parameters corresponding to group G_k. The proximal mapping for the $\ell_{1,2}$ group-norm penalty is

$$\text{prox}^{\ell_{1,2}}_{C_t+\delta I}(\theta_t) = \underset{\theta}{\text{argmin}} \left\{ \frac{1}{2}(\theta - \theta_t)^T (C_t + \delta I)(\theta - \theta_t) + \alpha_t \lambda \sum_{k \in K} \|\theta_{(k)}\|_2 \right\} \tag{31}$$

where C_t is a diagonal matrix.

The problem is separable with respect to the groups. For each group G_k we have to solve

$$\text{prox}^{\ell_{1,2}}_{D_{t,(k)}}(\theta_{t,(k)}) = \underset{\theta_{(k)}}{\text{argmin}} \left\{ \frac{1}{2}(\theta_{(k)} - \theta_{t,(k)})^T D_{t,(k)} (\theta_{(k)} - \theta_{t,(k)}) + \alpha_t \lambda \|\theta_{(k)}\|_2 \right\}, \tag{32}$$

where $D_{t,(k)} = C_{t,(k)} + \delta I_{(k)}$.

The solution is provided in the following Lemma.

Lemma 1. *The solution to (32) is given by*

$$\text{prox}^{\ell_{1,2}}_{D_{t,G_k}}(\theta_{t,G_k}) = \begin{cases} 0, & \|D_{t,(k)}\theta_{t,(k)}\|_2 \le \alpha_t \lambda, \\ \widetilde{D}_{t,(k)}\theta_{t,(k)}, & \|D_{t,(k)}\theta_{t,(k)}\|_2 > \alpha_t \lambda \end{cases}, \tag{33}$$

where $\widetilde{D}_{t,(k)}$ is a diagonal matrix with diagonal entries given by

$$[\widetilde{D}_{t,(k)}]_{ii} = \frac{[D_{t,(k)}]_{ii}}{[D_{t,(k)}]_{ii} + \alpha_t \lambda / \xi},$$

and ξ is defined as the unique solution to

$$1 = \sum_{i \in G_k} \left(\frac{[D_{t,(k)}]_{ii} [\theta_{t,(k)}]_i}{\xi [D_{t,(k)}]_{ii} + \alpha_t \lambda} \right)^2. \tag{34}$$

Proof. The proximal problem is a strongly convex quadratic that has a solution. If the solution is non-zero, the objective function is differentiable, and the solution satisfies

$$0 = D_{t,(k)}(\theta_{(k)} - \theta_{t,(k)}) + \alpha_t \lambda \frac{\theta_{(k)}}{\|\theta_{(k)}\|_2}$$

from which the form of $\widetilde{D}_{t,(k)}$ immediately follows. Plugging the above optimality condition into (32) results in a scalar problem in $\xi = \|\theta_{(k)}\|_2$:

$$\min_{\xi > 0} \xi + \frac{\theta_{t,(k)}^T \widetilde{D}_{t,(k)}(\xi) \theta_{t,(k)}}{\xi}. \tag{35}$$

When $\xi > 0$ we can differentiate and see that it satisfies (34). Since (35) is also convex, when (34) has a solution we know that this solution uniquely identifies the global minimizer ξ. The function $\sum_{i \in G_k} \left(\frac{[D_{t,(k)}]_{ii} [\theta_{t,(k)}]_i}{\xi [D_{t,(k)}]_{ii} + \alpha_t \lambda} \right)^2$ is monotonically decreasing in ξ. Hence (34) has a solution when the function at $\xi = 0$ is greater than 1, and this is equivalent to the condition in (33). □

In conclusion, to compute the proximal mapping for each group, we check the condition in (33) and find ξ using a root-finding method (e.g. bisection) observing that we have simple bounds for the root of (34): $0 < \xi < |G_k| \max_{i \in G_k}([D_{t,(k)}]_{ii})$, where $|G_k|$ is the cardinality of group G_k.

G Examples Satisfying Condition (C-4)

We provide concrete examples and derivations satisfying Condition (C-4) in Section 3. Before presenting our derivations, we need the following important theorem.

Theorem 2 (Weyl)**.** *For any two $n \times n$ Hermitian matrices A and B, assume that the eigenvalues of A and B are*

$$\mu_1 \ge \cdots \ge \mu_n, \quad \text{and} \quad \nu_1 \ge \cdots \ge \nu_n$$

respectively. Let $\lambda_1 \ge \cdots \ge \lambda_n$ be the eigenvalues of the matrix $A + B$, then the following holds

$$\mu_j + \nu_k \le \lambda_i \le \mu_r + \nu_s$$

for $j + k - n \ge i \ge r + s - 1$. Hence, we could derive

$$\lambda_1 \le \mu_1 + \nu_1$$

Algorithm 2 PROXGENW: A **Gen**eral Stochastic **Prox**imal Gradient Method with Weight Decay

Input: Stepsize α_t, $\{\rho_t\}_{t=1}^{t=T} \in [0,1)$, regularization parameter λ, small constant $0 < \delta \ll 1$, and weight decay regularization parameter ζ.
Initialize: $\theta_1 \in \mathbb{R}^d$, $m_0 = 0$, and $C_0 = 0$.
for $t = 1, 2, \ldots, T$ **do**
  Draw a minibatch sample ξ_t from $\mathbb{P}$
  $g_t \leftarrow \nabla f(\theta_t; \xi_t)$ ▷ Stochastic gradient at time t
  $m_t \leftarrow \rho_t m_{t-1} + (1-\rho_t) g_t$ ▷ First-order momentum estimate
  $C_t \leftarrow$ Preconditioner construction
  $\bar{\theta}_t \leftarrow (1 - \alpha_t \zeta)\theta_t$ ▷ Apply decoupled weight decay
  $\theta_{t+1} \in \underset{\theta \in \Omega}{\operatorname{argmin}} \Big\{ \langle m_t, \theta \rangle + \lambda \mathcal{R}(\theta) + \frac{1}{2\alpha_t} (\theta - \bar{\theta}_t)^\mathsf{T} \big(C_t + \delta I\big)(\theta - \bar{\theta}_t) \Big\}$
end for
Output: θ_T

Now, we derive the γ in (C-4) for various popular optimization algorithms used in deep learning communities.

ADAGRAD. In PROXGEN framework, ADAGRAD corresponds to $C_t = \Big(\frac{1}{t}\sum_{\tau=1}^{t} g_\tau g_\tau^\mathsf{T}\Big)^{1/2}$. Under the constant stepsizes $\alpha_t = \alpha$, we have

$$\begin{aligned}
\lambda_{\max}(C_t) &= \frac{1}{\sqrt{t}} \lambda_{\max}\Big(\sum_{\tau=1}^{t} g_\tau g_\tau^\mathsf{T}\Big)^{1/2} \\
&\leq \frac{1}{\sqrt{t}} \Big(\sum_{\tau=1}^{t} \lambda_{\max}(g_\tau g_\tau^\mathsf{T})\Big)^{1/2} \\
&= \frac{1}{\sqrt{t}} \Big(\sum_{\tau=1}^{t} \|g_\tau\|_2^2\Big)^{1/2} \\
&\leq G
\end{aligned}$$

Hence, the Condition (C-4) can be satisfied as

$$\lambda_{\min}(\alpha_t (C_t + \delta I)^{-1}) \geq \frac{\alpha}{G + \delta} \coloneqq \gamma$$

RMSPROP and ADAM. Exponential moving average (a.k.a. EMA) approaches correspond to $C_t = \big(\beta C_{t-1} + (1-\beta) g_t g_t^\mathsf{T}\big)^{1/2}$ where $\beta \in [0,1)$ and g_t denotes the stochastic gradient at time t. The usual RMSPROP and ADAM use diagonal approximations for $g_t g_t^\mathsf{T}$, but here we consider more general form (i.e. including general full matrix gradient outer-product) as introduce in [39]. First, we derive the upper bound for maximum eigenvalue for the matrix C_t. The matrix C_t can be expressed by

$$\begin{aligned}
C_t &= \big(\beta C_{t-1} + (1-\beta) g_t g_t^\mathsf{T}\big)^{1/2} \\
&= \big(\beta^2 C_{t-2} + \beta(1-\beta) g_{t-1} g_{t-1}^\mathsf{T} + (1-\beta) g_t g_t^\mathsf{T}\big)^{1/2} \\
&= \cdots \\
&= \Big((1-\beta) \sum_{i=1}^{t} \beta^{t-i} g_i g_i^\mathsf{T}\Big)^{1/2}
\end{aligned}$$

We can derive the upper bound by

$$\begin{aligned}
\lambda_{\max}(C_t) &= \lambda_{\max}\Big((1-\beta) \sum_{i=1}^{t} \beta^{t-i} g_i g_i^\mathsf{T}\Big)^{1/2} \\
&\leq \Big((1-\beta) \sum_{i=1}^{t} \beta^{t-i} \lambda_{\max}(g_i g_i^\mathsf{T})\Big)^{1/2} \\
&\leq \Big((1-\beta) G^2 \sum_{i=1}^{t} \beta^{t-i}\Big)^{1/2} \\
&\leq G(1-\beta^t)^{1/2} \leq G
\end{aligned}$$

Hence, we have $\lambda_{\max}(C_t + \delta I) \leq dG + \delta$. Also, we have

$$\lambda_{\max}\big(C_t + \delta I\big) = \frac{1}{\lambda_{\min}\big((C_t + \delta I)^{-1}\big)} \leq \frac{1}{G + \delta}$$

Therefore, the condition (C-4) under the constant stepsize $\alpha_t = \alpha$ can be derived as

$$\lambda_{\min}\big(\alpha_t (C_t + \delta I)^{-1}\big) \geq \frac{\alpha}{G + \delta}$$

which yields $\gamma = \frac{\alpha}{G+\delta}$.

Natural Gradient Descent. In this case, we derive the condition (C-4) for the Fisher information matrix when the loss function is defined as a negative log-likelihood, i.e., $f = \log p(x|\theta)$. The natural gradient descent aims at considering general geometry (not limited to Euclidean geometry), but we restrict our focus on the distribution space where the Fisher information is employed for preconditioner matrix C_t. The Fisher information matrix is defined as

$$F = \mathbb{E}_{Q(x)P(y|x,\theta)}\Big[\frac{\partial f(x|\theta)}{\partial \theta}\frac{\partial f(x|\theta)}{\partial \theta}^{\mathsf{T}}\Big]$$

where $Q(x)$ is data distribution and $P(y|x,\theta)$ denotes the model's predictive distribution (ex. neural networks). However, in general, we do not have access to true data distribution, so we instead take an expectation with respect to empirical (training) data distribution $\widehat{Q}(x)$. This trick is also employed for K-FAC approximations to the Fisher [36]. Let the training samples be $\mathcal{S} = \{x_1, \cdots, x_n\}$ with sample size n. Then, the empirical Fisher could be computed as

$$\begin{aligned}\widehat{F} &= \mathbb{E}_{\widehat{Q}(x)P(y|x,\theta)}\Big[\frac{\partial f(x|\theta)}{\partial \theta}\frac{\partial f(x|\theta)}{\partial \theta}^{\mathsf{T}}\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial f(x_i|\theta)}{\partial \theta}\frac{\partial f(x_i|\theta)}{\partial \theta}^{\mathsf{T}}\end{aligned}$$

Now, we bound the maximum eigenvalue of $\widehat{F}$ as

$$\begin{aligned}\lambda_{\max}(\widehat{F}) &\leq \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}\Big(\frac{\partial f(x_i|\theta)}{\partial \theta}\frac{\partial f(x_i|\theta)}{\partial \theta}^{\mathsf{T}}\Big) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} G^2 \\ &= G^2\end{aligned}$$

by our Condition (C-3). Hence, the Condition (C-4) can be derived as

$$\lambda_{\min}\big(\alpha_t (\widehat{F} + \delta I)^{-1}\big) \geq \frac{\alpha}{G^2 + \delta}$$

under the constant stepsize $\alpha_t = \alpha$.

H Proofs of Theorem 1

Lemma 2. *The first-order momentum m_t in Algorithm 1 satisfies*

$$\|m_t\|_2 \leq G$$

Proof. We use mathematical induction. For $t = 1$, the momentum is computed as $m_1 = \rho_1 m_0 + (1 - \rho_1) g_t = (1 - \rho_0) g_1$. Therefore, we have $\|m_t\|_2 = \|(1 - \rho_0) g_1\| \leq (1 - \rho_0) G \leq G$.

Now, we assume that $\|m_{t-1}\|_2 \leq G$ holds. The momentum at time t is constructed by $m_t = (1 - \rho_t) m_{t-1} + \rho_t g_t$. Then, we have

$$\begin{aligned}\|m_t\|_2 &= \|(1 - \rho_t) m_{t-1} + \rho_t g_t\|_2 \\ &\leq (1 - \rho_t)\|m_{t-1}\|_2 + \rho_t \|g_t\|_2 \\ &\leq (1 - \rho_t) G + \rho_t G = G\end{aligned}$$

where the first inequality comes from the triangle inequality and the second one is derived from the induction hypothesis. □

We deal with the following update rule in Algorithm 1 as

$$\theta_{t+1} \in \underset{\theta\in\Omega}{\operatorname{argmin}} \left\{ \langle (1-\rho_t)g_t + \rho_t m_{t-1}, \theta \rangle + \mathcal{R}(\theta) + \frac{1}{2\alpha_t}(\theta-\theta_t)^\top (C_t+\delta I)(\theta-\theta_t) \right\} \tag{36}$$

By the optimality condition, we have

$$0 \in (1-\rho_t)g_t + \rho_t m_{t-1} + \widehat{\partial}\mathcal{R}(\theta_{t+1}) + \frac{1}{\alpha_t}(C_t+\delta I)(\theta_{t+1}-\theta_t)$$

which means that

$$-(1-\rho_t)g_t - \rho_t m_{t-1} - \frac{1}{\alpha_t}(C_t+\delta I)(\theta_{t+1}-\theta_t) \in \widehat{\partial}\mathcal{R}(\theta_{t+1})$$

By adding the gradient $\nabla f(\theta_{t+1})$ on both sides, we have

$$\nabla f(\theta_{t+1}) - (1-\rho_t)g_t - \rho_t m_{t-1} - \frac{1}{\alpha_t}(C_t+\delta I)(\theta_{t+1}-\theta_t) \in \nabla f(\theta_{t+1}) + \widehat{\partial}\mathcal{R}(\theta_{t+1}) = \widehat{\partial}F(\theta_{t+1})$$

By the definition of θ_{t+1} in (36), we obtain

$$\begin{aligned} &\langle (1-\rho_t)g_t + \rho_t m_{t-1}, \theta_{t+1} \rangle + \mathcal{R}(\theta_{t+1}) + \frac{1}{2\alpha_t}(\theta_{t+1}-\theta_t)^\top (C_t+\delta I)(\theta_{t+1}-\theta_t) \\ \leq\ &\langle (1-\rho_t)g_t + \rho_t m_{t-1}, \theta_t \rangle + \mathcal{R}(\theta_t) \end{aligned}$$

which in result

$$\langle (1-\rho_t)g_t + \rho_t m_{t-1}, \theta_{t+1}-\theta_t \rangle + \mathcal{R}(\theta_{t+1}) + \frac{1}{2\alpha_t}(\theta_{t+1}-\theta_t)^\top (C_t+\delta I)(\theta_{t+1}-\theta_t) \leq \mathcal{R}(\theta_t)$$

Since the function f is L-smooth by Condition (C-1), we have

$$f(\theta_{t+1}) \leq f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1}-\theta_t \rangle + \frac{L}{2}\|\theta_{t+1}-\theta_t\|_2^2$$

Adding previous two inequalities yields

$$\begin{aligned} &\langle (1-\rho_t)g_t - \nabla f(\theta_t) + \rho_t m_{t-1}, \theta_{t+1}-\theta_t \rangle + (\theta_{t+1}-\theta_t)^\top \left(\frac{1}{2\alpha_t}(C_t+\delta I) - \frac{L}{2} I \right)(\theta_{t+1}-\theta_t) \\ \leq\ &F(\theta_t) - F(\theta_{t+1}) \end{aligned} \tag{37}$$

Then, we have

$$\begin{aligned} &\|\theta_{t+1}-\theta_t\|^2_{\frac{1}{2\alpha_t}(C_t+\delta I)-\frac{L}{2}I} \\ \overset{①}{\leq}\ &F(\theta_t) - F(\theta_{t+1}) - \langle (1-\rho_t)g_t - \nabla f(\theta_t), \theta_{t+1}-\theta_t \rangle - \langle \rho_t m_{t-1}, \theta_{t+1}-\theta_t \rangle \\ =\ &F(\theta_t) - F(\theta_{t+1}) - \langle g_t - \nabla f(\theta_t), \theta_{t+1}-\theta_t \rangle + \langle \rho_t g_t, \theta_{t+1}-\theta_t \rangle - \langle \rho_t m_{t-1}, \theta_{t+1}-\theta_t \rangle \\ \overset{②}{\leq}\ &F(\theta_t) - F(\theta_{t+1}) + \frac{1}{2L}\|g_t - \nabla f(\theta_t)\|_2^2 + \frac{L}{2}\|\theta_{t+1}-\theta_t\|_2^2 + \frac{\rho_t^2}{2L}\|g_t\|_2^2 + \frac{L}{2}\|\theta_{t+1}-\theta_t\|_2^2 \\ &+ \|\rho_t m_{t-1}\|_2 \|\theta_{t+1}-\theta_t\|_2 \\ \overset{③}{\leq}\ &F(\theta_t) - F(\theta_{t+1}) + \rho_0 \mu^{t-1} DG + \frac{\rho_0^2 \mu^{2(t-1)} G^2}{2L} + L\|\theta_{t+1}-\theta_t\|_2^2 + \frac{1}{2L}\|g_t - \nabla f(\theta_t)\|_2^2 \end{aligned}$$

The derivations in inequalities (1-3) as follows:

① We rearrange the inequality (37).

② We use the fact that $\langle a, b \rangle \leq \frac{1}{2}\|a\|_2^2 + \frac{1}{2}\|b\|_2^2$ and $\langle a, b \rangle \leq \|a\|_2 \|b\|_2$. With this, we use modified version such as $\langle a, b \rangle = \langle ca, \frac{1}{c}b \rangle \leq c^2\|a\|_2^2 + \frac{1}{c^2}\|b\|_2^2$ for any positive constant c.

③ We apply our Lemma 2 and Condition (C-3).

By rearranging the above inequality, we require the following quantity be positive-semidefinite.

$$\frac{1}{2\alpha_t}(C_t+\delta I) - \frac{3}{2}LI \succeq 0$$

Note that in this inequality we can see that

$$\frac{1}{2\alpha_t}(C_t+\delta I) - \frac{3}{2}LI \succeq \frac{1}{2\alpha_0}\delta I - \frac{3}{2}LI$$

since C_t is positive (semi)definite and α_t is *non-increasing*. Therefore, from this we can derive the stepsize condition in our Theorem 1 as

$$\alpha_0 \leq \frac{\delta}{3L}$$

Therefore, we have

$$\sum_{t=0}^{T-1} \|\theta_{t+1}-\theta_t\|^2_{\frac{1}{2\alpha_t}(C_t+\delta I)-\frac{3}{2}LI} \leq \underbrace{F(\theta_0)-F(\theta^*)}_{\Delta} + \underbrace{\frac{\rho_0 DG}{1-\mu}+\frac{\rho_0^2 G^2}{2L(1-\mu^2)}}_{C_1} + \frac{1}{2L}\sum_{t=0}^{T-1}\|g_t-\nabla f(\theta_t)\|_2^2$$

$$\leq \Delta + C_1 + \frac{1}{2L}\sum_{t=0}^{T-1}\|g_t-\nabla f(\theta_t)\|_2^2$$

Furthermore, we also have by stepsize condition

$$\Big(\frac{\delta}{2\alpha_0}-\frac{3}{2}L\Big)\sum_{t=0}^{T-1}\|\theta_{t+1}-\theta_t\|_2^2 \leq \sum_{t=0}^{T-1}\|\theta_{t+1}-\theta_t\|^2_{\frac{1}{2\alpha_t}(C_t+\delta I)-\frac{3}{2}LI} \leq \Delta + C_1 + \frac{1}{2L}\sum_{t=0}^{T-1}\|g_t-\nabla f(\theta_t)\|_2^2$$

since $\delta I \preceq C_t + \delta I$. From above inequality, we obtain

$$\sum_{t=0}^{T-1}\|\theta_{t+1}-\theta_t\|_2^2 \leq H_1 + H_2\sum_{t=0}^{T-1}\|g_t-\nabla f(\theta_t)\|_2^2 \tag{38}$$

where the constants H_1 and H_2 are defined as

$$H_1 = \Delta \Big/ \Big(\frac{\delta}{2\alpha_0}-\frac{3}{2}L\Big) + C_1 \Big/ \Big(\frac{\delta}{2\alpha_0}-\frac{3}{2}L\Big)$$

$$H_2 = \frac{1}{2L(\frac{\delta}{2\alpha_0}-\frac{3}{2}L)}$$

Our goal is to bound the distance between the zero vector and subdifferential set of F, so we have

$$\begin{aligned}
&\operatorname{dist}(\mathbf{0}, \widehat{\partial} F(\theta_{t+1}))^2 \\
=\ & \Big\|(1-\rho_t)g_t - \nabla f(\theta_{t+1}) + \rho_t m_{t-1} + \frac{1}{\alpha_t}(C_t+\delta I)(\theta_{t+1}-\theta_t)\Big\|_2^2 \\
=\ & \Big\|(1-\rho_t)g_t - \nabla f(\theta_{t+1}) + \rho_t m_{t-1} + (\theta_{t+1}-\theta_t) + \frac{1}{\alpha_t}(C_t+\delta I)(\theta_{t+1}-\theta_t) - (\theta_{t+1}-\theta_t)\Big\|_2^2 \\
\leq\ & 3\Big\|(1-\rho_t)g_t - \nabla f(\theta_{t+1}) + \rho_t m_{t-1} + (\theta_{t+1}-\theta_t)\Big\|_2^2 \\
&+ 3\Big\|\frac{1}{\alpha_t}(C_t+\delta I)(\theta_{t+1}-\theta_t)\Big\|_2^2 + 3\Big\|(\theta_{t+1}-\theta_t)\Big\|_2^2 \\
\leq\ & \underbrace{3\Big\|(1-\rho_t)g_t - \nabla f(\theta_{t+1}) + \rho_t m_{t-1} + (\theta_{t+1}-\theta_t)\Big\|_2^2}_{T_1} + 3\Big(\frac{1}{\gamma^2}+1\Big)\|\theta_{t+1}-\theta_t\|_2^2
\end{aligned}$$

Here, we assume that

$$\lambda_{\max}\big(\frac{1}{\alpha_t}(C_t+\delta I)\big) \leq \frac{1}{\gamma}$$

which yields our Condition (C-4)

$$\lambda_{\min}\big(\alpha_t(C_t+\delta I)^{-1}\big) \geq \gamma$$

From (37), we have

$$\big\langle (1-\rho_t)g_t - \nabla f(\theta_t) + \rho_t m_{t-1}, \theta_{t+1}-\theta_t\big\rangle + \big\|\theta_{t+1}-\theta_t\big\|^2_{\frac{1}{2\alpha_t}(C_t+\delta I)-\frac{L}{2}I} \leq F(\theta_t) - F(\theta_{t+1})$$

which can be re-written as

$$\begin{aligned}
&\Big\langle (1-\rho_t)g_t - \nabla f(\theta_{t+1}) + \rho_t m_{t-1}, \theta_{t+1}-\theta_t\Big\rangle \\
\leq\ & F(\theta_t) - F(\theta_{t+1}) - \big\langle \nabla f(\theta_{t+1}) - \nabla f(\theta_t), \theta_{t+1}-\theta_t\big\rangle - \big\|\theta_{t+1}-\theta_t\big\|^2_{\frac{1}{2\alpha_t}(C_t+\delta I)-\frac{L}{2}I} \\
\leq\ & F(\theta_t) - F(\theta_{t+1}) - \big\langle \nabla f(\theta_{t+1}) - \nabla f(\theta_t), \theta_{t+1}-\theta_t\big\rangle + \Big(\frac{\delta}{2\alpha_0}-\frac{L}{2}\Big)\|\theta_{t+1}-\theta_t\|_2^2
\end{aligned}$$

since we have the condition $\frac{\delta}{2\alpha_0} \geq \frac{3}{2}L$. Therefore, we obtain

$$\begin{aligned}
T_1 &= \|(1-\rho_t)g_t - \nabla f(\theta_{t+1}) + \rho_t m_{t-1}\|_2^2 + \|\theta_{t+1}-\theta_t\|_2^2 \\
&\quad + 2\Big\langle (1-\rho_t)g_t - \nabla f(\theta_{t+1}) + \rho_t m_{t-1}, \theta_{t+1}-\theta_t \Big\rangle \\
&\leq \|(1-\rho_t)g_t - \nabla f(\theta_t) + \nabla f(\theta_t) - \nabla f(\theta_{t+1}) + \rho_t m_{t-1}\|_2^2 + \|\theta_{t+1}-\theta_t\|_2^2 \\
&\quad + F(\theta_t) - F(\theta_{t+1}) - \langle \nabla f(\theta_{t+1}) - \nabla f(\theta_t), \theta_{t+1}-\theta_t\rangle + \Big(\frac{\delta}{2\alpha_0} - \frac{L}{2}\Big)\|\theta_{t+1}-\theta_t\|^2 \\
&\leq 4\|g_t - \nabla f(\theta_t)\|_2^2 + 4L^2\|\theta_{t+1}-\theta_t\|_2^2 + 4\|\rho_t m_{t-1}\|_2^2 + 4\|\rho_t g_t\|_2^2 + \|\theta_{t+1}-\theta_t\|_2^2 \\
&\quad + F(\theta_t) - F(\theta_{t+1}) + L\|\theta_{t+1}-\theta_t\|_2^2 + \Big(\frac{\delta}{2\alpha_0} - \frac{L}{2}\Big)\|\theta_{t+1}-\theta_t\|_2^2 \\
&\leq F(\theta_t) - F(\theta_{t+1}) + 4\rho_0^2\mu^{2(t-1)}G^2 + 4\rho_0^2\mu^{2(t-1)}G^2 \\
&\quad + \Big(\frac{\delta}{2\alpha_0} + \frac{L}{2} + 1 + 4L^2\Big)\|\theta_{t+1}-\theta_t\|_2^2 + 4\|g_t - \nabla f(\theta_t)\|_2^2
\end{aligned}$$

Therefore, we have the distance as

$$\begin{aligned}
&\mathrm{dist}\big(\mathbf{0}, \widehat{\partial} F(\theta_{t+1})\big)^2 \\
&\leq 3\Bigg(F(\theta_t) - F(\theta_{t+1}) + 8\rho_0^2\mu^{2(t-1)}G^2 + \underbrace{\Big(\frac{\delta}{2\alpha_0} + \frac{L}{2} + 2 + 4L^2 + \frac{1}{\gamma^2}\Big)}_{C_2}\|\theta_{t+1}-\theta_t\|_2^2 + 4\|g_t - \nabla f(\theta_t)\|_2^2\Bigg)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{dist}\big(\mathbf{0}, \widehat{\partial} F(\theta_a)\big)^2] &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|(1-\rho_t)g_t - \nabla f(\theta_{t+1}) + \rho_t m_{t-1} + \frac{1}{\alpha_t}(C_t + \delta I)(\theta_{t+1}-\theta_t)\Big\|_2^2\Big] \\
&\leq \frac{3}{T}\Big(\Delta + \frac{8\rho_0^2 G^2}{1-\mu^2} + 4\sum_{t=0}^{T-1}\|g_t - \nabla f(\theta_t)\|_2^2 + C_2\sum_{t=0}^{T-1}\|\theta_{t+1}-\theta_t\|_2^2\Big) \\
&\leq \frac{3}{T}\Big(\Delta + \frac{8\rho_0^2 G^2}{1-\mu^2} + 4\sum_{t=0}^{T-1}\|g_t - \nabla f(\theta_t)\|_2^2 + C_2(H_1 + H_2\sum_{t=0}^{T-1}\|g_t - \nabla f(\theta_t)\|_2^2\Big) \\
&\leq \frac{Q_1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|g_t - \nabla f(\theta_t)\|_2^2\big] + \frac{Q_2\Delta}{T} + \frac{Q_3}{T}
\end{aligned}$$

where

$$Q_1 = 4 + C_2H_2, \quad Q_2 = 3 + \frac{3C_2}{\frac{\delta}{2\alpha_0} - \frac{3}{2}L}, \quad Q_3 = \frac{24\rho_0^2G^2}{1-\mu^2} + \frac{3C_1C_2}{\frac{\delta}{2\alpha_0} - \frac{3}{2}L}$$

Note that the constants Q_1, Q_2, and Q_3 depend on $\{\alpha_0, \delta, L, D, G, \rho_0, \mu, \gamma\}$, but not on T. The third inequality comes from (38). If we assume the stochastic gradient g_t is evaluated on the minibatch $\mathcal{S}_t$ with $|\mathcal{S}_t| = b_t$, then we can obtain using Condition (C-2)

$$\begin{aligned}
\|g_t - \nabla f(\theta_t)\|_2^2 &= \mathbb{E}_\xi\Big[\Big\|\frac{1}{b_t}\sum_{i=1}^{b_t}\nabla f(\theta_t;\xi_{i_t}) - \nabla f(\theta_t)\Big\|_2^2\Big] \\
&= \frac{1}{b_t^2}\mathbb{E}\Big[\Big\|\sum_{i=1}^{b_t}\big\{\nabla f(\theta_t;\xi_{i_t}) - \nabla f(\theta_t)\big\}\Big\|_2^2\Big] \\
&\leq \frac{1}{b_t^2}\sum_{i_t=1}^{b_t}\mathbb{E}\big[\|\nabla f(\theta_t;\xi_{i_t}) - \nabla f(\theta_t)\|_2^2\big] \leq \frac{1}{b_t}\sigma^2
\end{aligned}$$

where i_t represents the random variable for each datapoint in minibatch samples $\mathcal{S}_t$. Finally, we arrive at our Theorem 1 as

$$\mathbb{E}_R[\mathrm{dist}\big(\mathbf{0}, \widehat{\partial} F(\theta_R)\big)^2] \leq \frac{Q_1\sigma^2}{T}\sum_{t=0}^{T-1}\frac{1}{b_t} + \frac{Q_2\Delta}{T} + \frac{Q_3}{T}$$

It can be clearly seen from the definitions of C_1, C_2, H_1, and H_2 that the constants $\{Q_i\}_{i=1}^3$ in Theorem 1 absolutely do not involve the problem dimension d.