Effects of momentum scaling for SGD

Dmitry A. Pasechnyuk DMIVILENSKY1@GMAIL.COM

Alexander Gasnikov GASNIKOV@YANDEX.RU

Martin Takáč MARTIN.TAKI@GMAIL.COM

Abstract

The paper studies the properties of stochastic gradient methods with preconditioning. We focus on momentum updated preconditioners with momentum coefficient β. Seeking to explain practical efficiency of scaled methods, we provide convergence analysis in a norm associated with preconditioner, and demonstrate that scaling allows one to get rid of gradients Lipschitz constant in convergence rates. Along the way, we emphasize important role of β, undeservedly set to constant $0.99...9$ at the arbitrariness of various authors. Finally, we propose the explicit constructive formulas for adaptive β and step size values.

1. Literature review

Preconditioning is long and widely known practice in numerical methods and mathematical programming [4–6]. With the recent surge of interest to statistical learning applications, there were proposed methods applicable to finite-sum function minimization [1, 7, 12]. In combination with momentum technique [14], preconditioning gave rise to adaptive methods, extensively used in applications [8, 11, 18, 21]. Recently, there has been a return to full-matrix methods using momentum [3, 10, 17, 20]. This gave significant benefit in practice, but in theory there still was no established with global convergence of SGD [2, 16]. Following [19], we analyse convergence of preconditioned gradient descent in preconditioner associated norm to get rid of gradient Lipschitz constant L. This allows us to determine step size independent on L and adaptive momentum parameter β. We also show the starting acceleration of convergence as in [9].

2. One-step effects

Firstly, we consider optimization problem $\min_{x\in\mathbb{R}^n} f(x)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is non-convex continuous function. Every additional requirement on f is introduced in appropriate place of the text where it is needed to simplify reasoning. Let us start with considering the simplest preconditioned method `Scaled SGD`

$$x_{t+1} = x_t - \eta_t P_t^{-1} g_t$$

with variable preconditioner $P_t \in \mathbb{S}^n_{++}$. Our goal for the nearest narration is to estimate the rate of function decreasing. For this purpose, we need to operate with a majorant of f. We assume that f has Lipschitz continuous gradient, that is $f(x_{t+1}) \leqslant f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L_t}{2}\|x_{t+1} - x_t\|^2$, on each of segment $[x_t, x_{t+1}]$, $t = 1, 2, ...$ of methods trajectory with a corresponding $L_t = L(x_t, x_{t+1})$. This assumption is weaker than uniform L-smoothness. This point of view allows to get rid of uniform constant L, but still does not tell anything about its value, because properties of the norm

$\|\cdot\|$ are not used anyhow. Then, if we go to norm $\|x\|_{P_t} = \langle P_t x, x\rangle^{1/2}$ associated with P_t, we have

$$f(x_{t+1}) \leqslant f(x_t) - \eta_t \langle \nabla f(x_t), P_t^{-1} g_t\rangle + \tfrac{L_t \eta_t^2}{2} \|g_t\|_{P_t}^{*2},$$

for some $L_t = L(x_t, x_{t+1}, P_t)$ which is believed to be smaller than previous $L(x_t, x_{t+1})$ if preconditioner is proper. The best case is when $P_t = \nabla^2 f(x_t)$ which implies that condition number is close to 1 in some vicinity of x_t, so it is also common to estimate convergence rate in $\|\cdot\|_{\nabla^2 f(x_t)}$ norm.

Another upper bound for f comes from M-Lipschitz continuity of $\nabla^2 f$:

$$f(x_{t+1}) \leqslant f(x_t) - \eta_t \langle \nabla f(x_t), P_t^{-1} g_t\rangle + \tfrac{\eta_t^2}{2} \langle \nabla^2 f(x_t) P_t^{-1} g_t, P_t^{-1} g_t\rangle + \tfrac{M\eta_t^3}{6} \|P_t^{-1} g_t\|^3,$$

which can be equivalently rewritten as

$$f(x_{t+1}) \leqslant f(x_t) - \eta_t \langle \nabla f(x_t), P_t^{-1} g_t\rangle + \tfrac{\eta_t^2}{2} \|g_t\|^{*2}_{P_t[\nabla^2 f(x_t)]^{-1} P_t^\top} + \tfrac{M\eta_t^3}{6} \|P_t^{-1} g_t\|^3.$$

Thus, we have got rid of L_t and replaced it with term without any uniform constant, but in different $P_t \left[\nabla^2 f(x_t)\right]^{-1} P_t^\top$ norm, plus additional cubic term. The new form of quadratic term gives us an opportunity to obtain the explicit replacement for L_t.

Indeed, we can think of $\nabla^2 f(x_t) P_t^{-1}$ as an inexactness of P_t, its closeness to the Hessian value, which can be bounded as follows:

$$\boxed{\nabla^2 f(x_t) P_t^{-1} \preccurlyeq (1 + \Delta_t) I}$$

where ideally $0 \leqslant \Delta_t \ll 1$. Hereinafter, $\preccurlyeq$ is used to compare arbitrary matrices with matrices of the form const $\cdot\, I$, so we can define it as follows: $A \preccurlyeq bI$ means $\lambda_{\max}(A) \leq b$. Then, we have

$$f(x_{t+1}) \leqslant f(x_t) - \eta_t \langle \nabla f(x_t), P_t^{-1} g_t\rangle + \tfrac{(1+\Delta_t)\eta_t^2}{2} \|g_t\|_{P_t}^{*2} + \tfrac{M\eta_t^3}{6} \|P_t^{-1} g_t\|^3.$$

Thus, we turned 2nd-order term with constant L into 2nd-order term with constant $(1 + \Delta_t)$, which is slightly greater than 1, and cubic term, so the behaviour of preconditioned gradient descent is as close to that of Newton method as P_t is to $\nabla^2 f(x_t)$.

From now, let us consider only preconditioners of the form

$$P_{t+1} = \beta_{t+1} P_t + (1 - \beta_{t+1}) d_{t+1}, \tag{1}$$

where $d_{t+1} = \operatorname{diag}\left(\nabla^2 f(x_{t+1})\right)$ and $P_0 = I$. One can use any other proper update instead of d_{t+1}, which preserves positive definiteness of P_t (if f is non-convex, positive truncation should be applied to d_{t+1}, see [13]). We assume that d_t is a good approximation of Hessian, so that P_t is maintained to be close to Hessian. In the case of diagonal d_t, we assume that $\nabla^2 f$ is almost diagonal. By introducing two more inexactness relating preconditioner and Hessian to update term d_t

$$\boxed{\nabla^2 f(x_t) d_t^{-1} \preccurlyeq (1 + \sigma) I} \qquad \boxed{(1 - \delta_t^-) I \preccurlyeq P_t d_t^{-1} \preccurlyeq (1 + \delta_t^+) I,}$$

we get the opportunity to express each one of σ, δ and Δ through other ones. Estimating a local Lipschitz constant of f after scaling, we obtain the following proposition that bound Δ.

Proposition 1 *For preconditioner updated in accordance with (1), inexactness Δ_t depends on inexactness δ_t^- as follows $\Delta_t \leqslant \frac{1+\sigma}{1-\min\{\delta_t^-, \beta_t\}} - 1$.*

Note, that the dependency on β_t is hidden behind δ_t^-. But δ_t^- grows with β_t, $\delta_t^- = 0$ for $\beta_t = 0$ and $\delta_t^- \in [0, 1)$, so our new bound on Δ_t behaves similarly to the previous one.
It is obvious that $\beta_t = 0$ is the best choice for the case $g_t = \nabla f(x_t)$. Otherwise, small β_t also leads to additional penalty on the variation of P_t. If g_t is unbiased estimator of $\nabla f(x_t)$, this penalty goes with variation $\|s\|_{P_t}^{*2}$, where $s = g_t - \nabla f(x_t)$. Since we consider $\|\cdot\|_{P_t}$ norm, penalty appears when we go from $\|\cdot\|_{P_t}$ to $\|\cdot\|_{P_{t+1}}$ norm.

Proposition 2 *For any $s \in \mathbb{R}^n$, it holds that* $\|s\|^{*2}_{P_{t+1}} \leqslant (1 + \frac{1-\beta_{t+1}}{1/\delta_t^+ + \beta_{t+1}})\|s\|^{*2}_{P_t}$.

Appearing factor is a penalty. For fixed δ_t^+, it decreases inversely proportional to β_t, have maximum in $\beta_t = 0$ with value $1 + \delta_t^+$ and minimum in $\beta_t = 1$ with value 1. If δ_t^+ depends on β_t, penalty may behave in a more complicated way, but still pushes the best value of β_t away from zero.
It remains to relate δ_t and β_t to obtain a descent lemma depending only on a choice of β_t. There are two ways to do this: assuming, that Hessian changes only a little from iteration to iteration, or not.

Proposition 3 *For any $(d_t \succcurlyeq 0)_{t=0}^{\infty}$, it holds that $\delta^+_{t+1} = \beta_{t+1}\kappa_{t+1}, \delta^-_{t+1} = \beta_{t+1}\chi_{t+1}$, where* $\kappa_t = \left[\frac{\max_i [P_{t-1}]_{ii}}{\min_i [d_t]_{ii}} - 1\right]_+, \chi_t = \left[1 - \frac{\min_i [P_{t-1}]_{ii}}{\max_i [d_t]_{ii}}\right]_+$.

Proposition 4 *If f is strong self-concordant [15], that is $\forall x, y, z, w \in \mathbb{R}^n$*

$$diag\left(\nabla^2 f(y) - \nabla^2 f(x)\right) \preccurlyeq N\|y - x\|_{diag(\nabla^2 f(z))} diag\left(\nabla^2 f(w)\right)$$

for some $N > 0$, then it holds that $\delta^+_{t+1} \leqslant \beta_{t+1}[\delta_t^+ + \delta_t^+\sqrt{1 + \delta_t^+}N\eta_t\|g_t\|^*_{P_t} - 1]_+$.

We can finally estimate the descent on the one step, depending only on β and not on any of inexactnesses.

Theorem 5 (Descent Lemma) *Point x_{t+1} generated by* `Scaled SGD` *on iteration t satisfies*

$$\begin{aligned}\mathbb{E}\left[f(x_{t+1})\right] \leqslant f(x_t) - \tfrac{\eta_t}{2}\|\nabla f(x_t)\|^{*2}_{P_t} + \tfrac{\eta_t}{2}((\tfrac{\eta_t(1+\sigma)}{1-\beta_t\chi_t})^2 + \tfrac{\eta_t(1+\sigma)}{1-\beta_t\chi_t} - 1)\|g_t\|^{*2}_{P_t} \\ + \tfrac{\eta_t}{2}(1 + \tfrac{(1-\beta_{t+1})\beta_t}{1/\kappa_t + \beta_t\beta_{t+1}})\|g_t - \nabla f(x_t)\|^{*2}_{P_t} + \tfrac{M'\eta_t^3}{6}(\tfrac{1+\sigma}{1-\beta_t\chi_t})^{3/2}\|g_t\|^{*3}_{P_t}.\end{aligned}$$

3. Cumulative effects

Further, we consider finite-sum optimization problem $\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{m}\sum_{i=1}^m f_i(x)$. Such a form of objective function became widely used due to the recent surge of interest in statistical learning applications.

We assume that Lipschitz constants are close to 1, in view of what we have demonstrated, but further we do not specify their values. This allows us to claim that following results stay valid not only for the preconditioned methods, but for any gradient method if the difference between L_t, $t = 1, 2, ...$ is significant.

Theorem 6 *If $\eta_t \leqslant \min\{\frac{\alpha p}{3}, \frac{3}{4}\frac{p}{5p+1}\}\frac{1}{L_t}$, sequence of the points generated by* `Scaled L-SVRG` *satisfies* $\mathbb{E}\left[\|\nabla f(\overline{x}_T)\|^{*2}_{P_t}\right] \leqslant \frac{4}{\alpha p}\frac{\overline{L}}{T}[f(x_0) - f(x_*) + 2\Gamma\sum_{t=2}^{T+1} L_t\mathbb{E}\left[(1-\beta_t)\|x_t - y_t\|_2^2\right]]$, *where* $\overline{L} = \frac{T}{\frac{1}{L_1} + \cdots + \frac{1}{L_T}}$ *is a harmonic average of L_t, $t = 1, ..., T$.*

Note, that harmonic average, which appears in convergence rate estimates too rarely by the way, has the wonderful property that it is minimum-dominated. In particular, $\frac{T}{\sum_{t=1}^T 1/L_t} \leqslant T\min_{t=1,...,T} L_t$.

This implies that it does not matter how big is one of the L_t (or all of them, except one), their harmonic average will be bounded and stuck to minimum of the elements being averaged, even if some of those variable L_t are infinite. Moreover, harmonic average is always less than arithmetic mean. So, we say that local Lipschitz constants are very well-aggregated in final convergence rate. If algorithm maintain L_t close to one for each $t = 1, 2, ...$, final convergence rate is almost independent on global Lipschitz constant!

Algorithm 1 `Scaled L-SVRG`

Data: $p \in (0,1), \{\eta_t > 0\}_{t=0}^{\infty}, \{0 < \beta_t < 1\}_{t=0}^{\infty}, x_0$
$P_0 = I, y_0 = x_0$
for $t \geqslant 0$ **do**
  Draw $i_t \in \{1, ..., m\}$ from uniform distribution
  $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(y_t) + \nabla f(y_t)$
  $y_{t+1} = \begin{cases} y_t & \text{with probability } p \\ x_t & \text{with probability } 1-p \end{cases}$
  $x_{t+1} = x_t - \eta_t P_t^{-1} g_t$
  Draw $z_t \in \{-1, 1\}^n$ from Rademacher distribution
  $P_{t+1} = \beta_{t+1} P_t + (1-\beta_{t+1})\text{diag}\left(z_t \circ \nabla^2 f(x_{t+1}) z_t\right)$
end for

Note that obtained dependence of the norm of gradient after T iterations of algorithm on β_t allows one to choose β_t so that error term in Theorem 6 can take a small value. At least, we should guarantee its boundedness for every T. This can be achieved by the several ways: by choosing β_t depending on pre-known number of iterations T, or by making it dependent on some hyperparameter sequence. In the first case, we notice that for $\beta_t = 1 - 1/(TL_t\|x_t - y_t\|_2^2)$, error term is equal to 2Γ and does not affect the convergence rate. But with this approach β_t values can be chosen too close to 1, which is not efficient in practice. To prevent this, we can bound the error term with the series with upper limit $+\infty$. Then, it is sufficient to choose $\beta_t = 1 - a_t/(L_t\|x_t - y_t\|_2^2)$, where $a_t > 0$ and $\sum_{t=2}^{+\infty} a_t$ is converging, so that error term is bounded.

4. Synthesis

In this section, we continue the convergence analysis started in "One-step effects" and based on descent lemma, using the sketch of the proof from "Cumulative effects". To simplify the reasoning, we consider `Scaled SGD` without variance reduction but updating step direction as $g_t = \nabla f(x_t)$ with probability $1 - p$.

The following corollary of Theorem 5 shows the convergence of `Scaled SGD` updating step direction as $g_t = \nabla f(x_t)$ with probability $1 - p$.

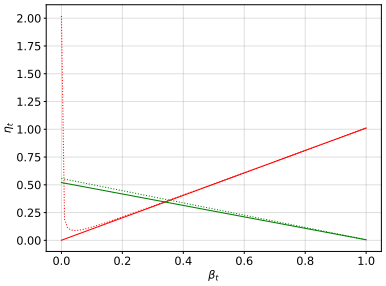

Figure 1: Intuition for the β_t choice.

Theorem 7 *If* $\eta_t \leqslant \min\left\{\frac{1-\beta_t\chi_t}{(1+\sigma)(\Phi+\sqrt{M'/6\cdot\|g_t\|_{P_t}^*})}, \frac{1/(1+\kappa_t)+\beta_t}{1-p}\right\}$, *sequence of the points generated by* `Scaled SGD` *satisfies* $\min_{t=1,\dots,T}\|\nabla f(x_t)\|_{P_t}^{*2} = O(\frac{f(x_0)-f(x_*)}{T})$.

Note that the first term in $\eta_t \leqslant \min\{\cdot,\cdot\}$ is decreasing, and the second one is increasing, so there is β_t at which upper bound on η_t has a fracture and starting from which it is determined by decreasing second term (see Figure 1). Thus, upper bound on η_t attains its maximum with respect to β_t at this point. So, we have equation $\frac{1-\beta_t\chi_t}{(1+\sigma)(\Phi+\sqrt{M'/6\cdot\|g_t\|_{P_t}^*})} = \frac{1/(1+\kappa_t)+\beta_t}{1-p}$, which leads to

$$\beta_t = \max\left\{\frac{\beta_{t-1}}{2}, \frac{(1-p)(1+\kappa_t+\chi_t)}{(1+\sigma)(\Phi+\sqrt{M'/6\cdot\|g_t\|_{P_t}^*})+(1-p)\chi_t} - 1\right\}. \tag{2}$$

5. Observed effects

Let's consider the binary logistic regression task $\min_{x\in\mathbb{R}^n}\{f(x) := \frac{1}{m}\sum_{i=1}^m \log(1+e^{-b_i(a\circ a_i)^\top x})\}$, where $\{(a_i, b_i)\}_{i=1}^m$ is a dataset containing features a_i and classes $b_i \in \{-1, 1\}$, and $a \in \mathbb{R}^n$ is vector of random i.i.d. scaling factors drawn from $\mathcal{U}[-A, A]$ and corresponding to each feature, $\circ$ denotes

Hadamard product. We use `a9a` dataset with $m = 32561$, $n = 123$. For logistic regression problem in this formulation, ∇f is Lipschitz continuous with constant $L = \frac{1}{4}\|(a \circ a_1 \ \ldots \ a \circ a_m)\|_2 = O(A)$, so that L is proportional to A, which is helpful for the design of experiments. In all the experiments we froze the following hyperparameters: batch size $= 100$, $p = 0.9$.

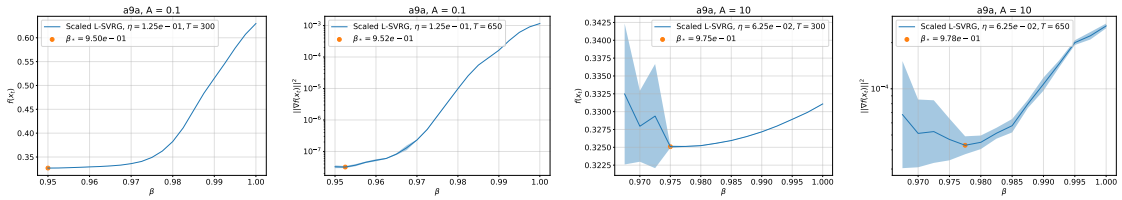

Figure 2: Dependence of achieved precision on $\beta_t \equiv \beta$.

Figure 2 summarizes the results of `Scaled L-SVRG` runs with different β's. Horizontal axis measures value of β, vertical axis measures objective function value $f(x_T)$ (or $\|\nabla f(x_T)\|^2$) after $T = 300$ iterations of the algorithm. Firstly, scaling give significant benefit in comparison with not scaled method, corresponding to $\beta = 1$. (More detailed experiments are presented in Appendices C and D). It can be seen that dependence of achieved precision on β changes with increasing of A: minimum of the corresponding function is getting closer to $\beta = 1$, its values on the left from minimum are growing and its growth rate near $\beta = 1$ is significantly increasing. This relationship between β and A reflects the trade-off between variance compensation and scaling. Variance affects the convergence if β is small: increasing of L leads to the increasing of δ_t^+, which increases the accumulating error term in Proposition 2. To explain the behaviour near $\beta = 1$, consider the $A = 0.1$ case, where variance error terms are insignificant. Values begin to grow rapidly starting from $\beta \approx 0.97$ and stop on some fixed value at $\beta = 1$. This behaviour is described in Proposition 1, where we have shown the $O(1/(1-\beta))$ growth of gradients Lipschitz constant. The boundedness at $\beta = 1$ can be explained by the proper choice of P_0, such that $\delta_t^- \neq 1$, even if $\beta = 1$. Thus, the main outline of our theory is successfully confirmed on the experiment.

The dependence of optimal β_* on smoothness characteristic A is presented on the Figure 3a. It can be seen that β_* grow slowly with increasing of A (plot is close to linear in logarithmic scale for A). This means that there is no need for β to be in proportional dependence with η or L. On the other hand, best choice of β is close to standard $\beta \approx 0.99$ for large enough L (and it does not matter, $L = 100$ or $L = 200$, because they are of the same order), while small L make choice of β very sensitive (that is, one need to make a distinctions between $L = 0.01$ and $L = 0.001$, although they are pretty close to each other).

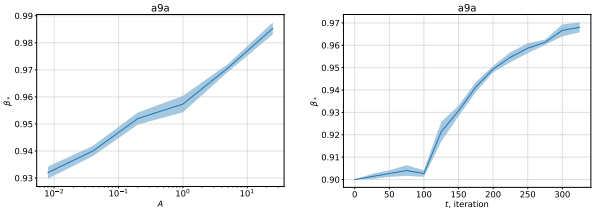

Figure 3: Dependencies of optimal β_*.

In addition to dependence of optimal β_* on smoothness of the problem, we test the dependence on the number of iterations. This experiment gives a rough estimate for the best choice of β_t. On the Figure 3b, one can see the dependence of quasi-optimal (in a sense described above) β_* on the number of iterations. Optimal β_* is getting closer to $\beta = 1$ with increasing number of iterations, while at the beginning of methods operation the best value is significantly lower. The latter can be explained by the need to rapidly adapt P_t to some good estimation of component-wise scaling from its initial value $P_0 = I$. On the other hand, the convergence of optimal β_* to 1 as the iteration number tends to infinity can be explained in view of remark from the end of "Cumulative effects" section.

References

[1] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276, 1998.

[2] Albert S Berahas, Majid Jahani, Peter Richtárik, and Martin Takáč. Quasi-newton methods for machine learning: forget the past, just sample. *Optimization Methods and Software*, pages 1–37, 2021.

[3] Aleksandr Beznosikov, Aibek Alanov, Dmitry Kovalev, Martin Takáč, and Alexander Gasnikov. On scaled methods for saddle point problems. *arXiv preprint arXiv:2206.08303*, 2022.

[4] Charles G Broyden. Quasi-newton methods and their application to function minimisation. *Mathematics of Computation*, 21(99):368–381, 1967.

[5] WC Davidon. Variable metric method for minimization (research and development report anl-5990, rev. ed.). *Argonne IL: Argonne National Laboratory, US Atomic Energy Commission*, 1959.

[6] John E Dennis and Jorge J Moré. Quasi-newton methods, motivation and theory. *SIAM review*, 19(1):46–89, 1977.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[8] Alina Ene and Huy Lê Nguyen. Adaptive and universal algorithms for variational inequalities with optimal convergence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (6):6559–6567, 2022.

[9] Slavomír Hanzely, Dmitry Kamzolov, Dmitry Pasechnyuk, Alexander Gasnikov, Peter Richtárik, and Martin Takáč. A damped Newton method achieves global $O\left(\frac{1}{k^2}\right)$ and local quadratic convergence rate. *Advances in Neural Information Processing Systems*, 35, 2022.

[10] Majid Jahani, Sergey Rusakov, Zheng Shi, Peter Richtárik, Michael W Mahoney, and Martin Takáč. Doubly adaptive scaled algorithm for machine learning using second-order information. *International Conference on Learning Representations (ICLR)*, 2021.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Xi-Lin Li. Preconditioner on matrix Lie group for SGD. *arXiv preprint arXiv:1809.10232*, 2018.

[13] Santiago Paternain, Aryan Mokhtari, and Alejandro Ribeiro. A Newton-based method for nonconvex optimization with fast evasion of saddle points. *SIAM Journal on Optimization*, 29 (1):343–368, 2019.

[14] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[15] Anton Rodomanov and Yurii Nesterov. Greedy quasi-Newton methods with explicit superlinear convergence. *SIAM Journal on Optimization*, 31(1):785–811, 2021.

[16] Anton Rodomanov and Yurii Nesterov. Rates of superlinear convergence for classical quasi-newton methods. *Mathematical Programming*, 194(1):159–190, 2022.

[17] Abdurakhmon Sadiev, Aleksandr Beznosikov, Abdulla Jasem Almansoori, Dmitry Kamzolov, Rachael Tappenden, and Martin Takáč. Stochastic gradient methods with preconditioned updates. *arXiv preprint arXiv:2206.00285*, 2022.

[18] Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates. In *International conference on machine learning*, pages 343–351. PMLR, 2013.

[19] Katya Scheinberg and Xiaocheng Tang. Practical inexact proximal quasi-newton method with global complexity analysis. *Mathematical Programming*, 160(1):495–529, 2016.

[20] Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. Adahessian: An adaptive second order optimizer for machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10665–10673, 2021.

[21] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Appendix A. Additional theory

Remark 8 *It turns out that formula in Theorem 5 includes both β_t and β_{t+1}. To find optimal β_t we should consider sum of descents on two corresponding steps, but it could make the analysis more complicated. One can assume $\beta_1 = \beta_2 = ... \equiv \beta$ to find optimal β. But it is easy to see that*

$$\arg\min_{\beta} \frac{(1-\beta)\beta}{1/\kappa_t + \beta^2} = 0, \quad \arg\min_{\beta} \frac{1+\sigma}{1-\beta_t\chi_t} = 0,$$

so we lose adaptivity. This defect is common for δ_t linearly dependent on β.

In practice, β is usually chosen to be close to 1, assuming that dependence of L_t on β is "weaker" than dependence of penalty for variance on β. So, it worth estimate the worst case (with respect to the β) multiplier appearing in variance term. This can be done explicitly

$$1 + \max_{\beta} \frac{(1-\beta)\beta}{1/\kappa_t + \beta^2} = 1 + \frac{1}{2}\left(\sqrt{\kappa_t + 1} - 1\right) = O\left(\sqrt{\frac{L}{\mu}}\right).$$

On the other hand, the worst case smoothness multiplier is L and is not being accumulated with iterations of the algorithm. Therefore, it was right to assume that β should be close to 1 giving priority to restrain the growth of variance term.

Remark 9 *Note, that second term in (2) is less than zero while $\|g_t\|^*_{P_t} \geqslant \frac{6}{M'}\left(\frac{(1-p)(1+\kappa_t)}{1+\sigma} - \Phi\right)^2$, so β_t should be set to zero on the first iterations (the more ill-conditioned is the function, the more iterations are needed). On the other hand, $\|g_t\|^*_{P_t} \to 0$, so β_t on the later iteration is determined by $\frac{(1-p)(1+\kappa_t) - (1+\sigma)\Phi}{(1-p)\chi_t + (1+\sigma)\Phi}$, which is smaller than one if $1 + \kappa_t - \chi_t < \frac{2\Phi(1+\sigma)}{1-p}$ (the more*

well-conditioned is the function, the less β_t is). In addition, note that β_t is increasing function of κ_t and decreasing function of χ_t.

Let's now establish the starting acceleration of the convergence, achieved by scaled methods. For this purpose we change (4) as

$$x \geqslant \frac{-1 + \sqrt{3 + 2A\|g_t\|^*_{P_t}}}{2(1 + A\|g_t\|^*_{P_t})},$$

which is solution to $(1 + A\|g_t\|^*_{P_t})x^2 + x - 1/2 = 0$. Then, following the analysis proposed in [9], we state

$$\eta_t \geqslant \frac{1}{4\max\left\{1, \sqrt{A\|g_t\|^*_{P_t}}\right\}},$$

which leads to

$$\eta_t\|g_t\|^*_{P_t} \geqslant \begin{cases} \dfrac{\|g_t\|^{*2}_{P_t}}{4}, & \text{if } \|g_t\|^*_{P_t} < 1/A \\ \dfrac{\|g_t\|^{*3/2}_{P_t}}{4\sqrt{A}}, & \text{otherwise.} \end{cases}$$

Thus, we have two cases of convergence lemma for Lyapunov function V_t introduced above:

$$\begin{cases} \mathbb{E}\left[V_{t+1}\right] \leqslant V_t - \dfrac{1}{16}(2\|\nabla f(x_t)\|^{*2}_{P_t} + \|g_t\|^{*2}_{P_t}), & \text{if } \|g_t\|^*_{P_t} < 1/A \\ \mathbb{E}\left[V_{t+1}\right] \leqslant V_t - \dfrac{1}{16\sqrt{A}}\left(2\dfrac{\|\nabla f(x_t)\|^{*2}_{P_t}}{\|g_t\|^{*1/2}_{P_t}} + \|g_t\|^{*3/2}_{P_t}\right), & \text{otherwise.} \end{cases}$$

Thus, at the starting iterations, when $\|g_t\|^*_{P_t} \geqslant \dfrac{3}{M'}\sqrt{\dfrac{1+\sigma}{1-\beta_t\chi_t}}$, convergence rate of scaled method is $O\left(\dfrac{1}{T^{3/2}}\right)$. Moreover, the greater is β_t, the shorter this starting acceleration lasts. But $\min\limits_{\beta_t} A = \dfrac{M'}{3\sqrt{1+\sigma}}$, so there cannot be an acceleration on the latter iterations. Anyway, β_t could be chosen in a way to lengthen this starting convergence.

Appendix B. Omitted proofs

B.1. Proof of Proposition 1

Then, inexactness Δ_{t+1} can be estimated as follows

$$\nabla^2 f(x_{t+1})P^{-1}_{t+1} = \frac{1}{1-\beta_{t+1}}\nabla^2 f(x_{t+1})d^{-1}_{t+1} - \frac{1}{1-\beta_{t+1}}\nabla^2 f(x_{t+1})d^{-1}_{t+1}\left[\frac{1-\beta_{t+1}}{\beta_{t+1}}d_{t+1}P^{-1}_t + I\right]^{-1},$$

using Woodbury identity

$$(A+B)^{-1} = A^{-1} - A^{-1}(AB^{-1} + I)^{-1},$$

that implies

$$\Delta_t \leqslant \frac{1+\sigma}{1-\beta_t} - 1,$$

which is increasing with β, always greater than σ, increases linearly near $\beta = 0$ and inversely proportional near $\beta = 1$.

Estimation for Δ is rough near $\beta = 1$. Its because we neglect second term in Woodbury identity. To work it out, we use the relation between P_t and d_{t+1} to bound the second term in Woodbury identity, using the formula for P_{t+1}:

$$\frac{1}{1-\beta_{t+1}} P_{t+1} d_{t+1}^{-1} = \frac{\beta_{t+1}}{1-\beta_{t+1}} P_t d_{t+1}^{-1} + I,$$

$$\frac{\beta_{t+1}}{1-\beta_{t+1}} P_t d_{t+1}^{-1} = \frac{1}{1-\beta_{t+1}} \left(P_{t+1} d_{t+1}^{-1} - I + \beta_{t+1} I \right) \succcurlyeq \frac{\beta_{t+1} - \delta_{t+1}^-}{1-\beta_{t+1}} I,$$

$$\frac{1-\beta_{t+1}}{\beta_{t+1}} d_{t+1} P_t^{-1} + I \preccurlyeq \frac{1-\delta_{t+1}^-}{\beta_{t+1} - \delta_{t+1}^-} I,$$

$$\left[\frac{1-\beta_{t+1}}{\beta_{t+1}} d_{t+1} P_t^{-1} + I \right]^{-1} \succcurlyeq \frac{\beta_{t+1} - \delta_{t+1}^-}{1-\delta_{t+1}^-} I.$$

Substituting this bound in Woodbury identity finishes the proof.

B.2. Proof of Proposition 2

Let's apply Woodbury identity to P_{t+1}:

$$P_{t+1}^{-1} = \frac{1}{\beta_{t+1}} P_t^{-1} - \frac{1}{\beta_{t+1}} P_t^{-1} \left[\frac{\beta_{t+1}}{1-\beta_{t+1}} P_t d_t^{-1} + I \right]^{-1}.$$

Then,

$$\langle s, P_{t+1}^{-1} s \rangle = \frac{1}{\beta_{t+1}} \langle s, P_t^{-1} s \rangle - \frac{1}{\beta_{t+1}} \langle s, \left(\left[\frac{\beta_{t+1}}{1-\beta_{t+1}} P_t d_t^{-1} + I \right] P_t \right)^{-1} s \rangle.$$

On the other hand,

$$\frac{\beta_{t+1}}{1-\beta_{t+1}} P_t d_t^{-1} + I \preccurlyeq \left(\frac{\beta_{t+1}}{1-\beta_{t+1}} (1 + \delta_t^+) + 1 \right) I \preccurlyeq \frac{1 + \beta_{t+1} \delta_t^+}{1-\beta_{t+1}} I.$$

Finally, we substitute this bound in equality on variation and obtain that

$$\|s\|_{P_{t+1}}^{*2} \leqslant \frac{1+\delta_t^+}{1+\beta_{t+1}\delta_t^+} \|s\|_{P_t}^{*2} = \left(1 + \frac{1-\beta_{t+1}}{1/\delta_t^+ + \beta_{t+1}} \right) \|s\|_{P_t}^{*2}. \tag{3}$$

B.3. Proof of Proposition 3

It follows from

$$P_{t+1} d_{t+1}^{-1} = (1-\beta_{t+1}) I + \beta_{t+1} P_t d_{t+1}^{-1} \preccurlyeq \left[1 + \beta_{t+1} \left(\frac{\max_i [P_t]_{ii}}{\min_i [d_{t+1}]_{ii}} - 1 \right) \right] I.$$

B.4. Proof of Proposition 4

We need in the following corollary of strong self-concordance:

Lemma 10 (Rodomanov–Nesterov [15]) *For all $x, y \in \mathbb{R}^n$, it holds that*

$$\frac{\mathit{diag}\left(\nabla^2 f(x)\right)}{1 + N\|y-x\|_{\mathit{diag}(\nabla^2 f(x))}} \preccurlyeq \mathit{diag}\left(\nabla^2 f(y)\right) \preccurlyeq \left(1 + N\|y-x\|_{\mathit{diag}(\nabla^2 f(x))}\right) \mathit{diag}\left(\nabla^2 f(x)\right),$$

where $A \preccurlyeq B$ means $\langle (B-A)x, x \rangle \geqslant 0$ for all $x \in \mathbb{E}$.

Presented lemma implies the bound on δ which takes into account that $P_t d_{t+1}^{-1}$ term is as small as the step on iteration t. We have

$$\left(1 + N\eta_t \|g_t\|_{P_t}^* \sqrt{1+\delta_t^+}\right)^{-1} d_t \preccurlyeq d_{t+1} \preccurlyeq (1 + N\|x_{t+1} - x_t\|_{d_t})\, d_t \preccurlyeq \left(1 + N\eta_t \|g_t\|_{P_t}^* \sqrt{1+\delta_t^+}\right) d_t,$$

$$P_{t+1} d_{t+1}^{-1} = (1-\beta_{t+1})I + \beta_{t+1} P_t d_{t+1}^{-1} \preccurlyeq \left[1 + \beta_{t+1}(\delta_t^+ + \delta_t^+\sqrt{1+\delta_t^+} N\eta_t \|g_t\|_{P_t}^* - 1)\right] I,$$

which implies the statement of the proposition.

B.5. Proof of Theorem 5

$$f(x_{t+1}) \leqslant f(x_t) - \eta_t \langle \nabla f(x_t), P_t^{-1} g_t\rangle + \frac{\eta_t^2}{2}\frac{1+\sigma}{1-\beta_t\chi_t}\|g_t\|_{P_t}^{*2} + \frac{M'\eta_t^3}{6}\left(\frac{1+\sigma}{1-\beta_t\chi_t}\right)^{3/2}\|g_t\|_{P_t}^{*3}$$

$$-\eta_t\langle \nabla f(x_t), P_t^{-1} g_t\rangle = -\frac{\eta_t}{2}\|\nabla f(x_t)\|_{P_t}^{*2} - \frac{\eta_t}{2}\|g_t\|_{P_t}^{*2} + \frac{\eta_t}{2}\|g_t - \nabla f(x_t)\|_{P_t}^{*2}$$

$$\mathbb{E}\left[\|g_{t+1} - \nabla f(x_{t+1})\|_{P_{t+1}}^{*2}\right] \leqslant \left(1 + \frac{(1-\beta_{t+1})\beta_t}{1/\kappa_t + \beta_t\beta_{t+1}}\right)\|g_t - \nabla f(x_t)\|_{P_t}^{*2} + \mathbb{E}\left[\|\nabla f(x_{t+1}) - \nabla f(x_t)\|_{P_t}^{*2}\right]$$

$$\|\nabla f(x_{t+1}) - \nabla f(x_t)\|_{P_t}^{*2} \leqslant \eta_t^2\left(\frac{1+\sigma}{1-\beta_t\chi_t}\right)^2\|g_t\|_{P_t}^{*2}$$

B.6. Proof of Theorem 6

We will demonstrate the convergence of the method for such a Lyapunov function $V_t = f(x_t) - f(x_*) + a_t\|x_t - y_t\|_{P_t}^2$. Due to L_t-Lipschitz smoothness of f, we have

$$\mathbb{E}\left[V_{t+1}\right] \leqslant f(x_t) - f(x_*) - \eta_t\langle \nabla f(x_t), P_t^{-1}\nabla f(x_t)\rangle + \frac{\eta_t^2 L_t}{2}\mathbb{E}\left[\|g_t\|_{P_t}^{*2}\right] + a_{t+1}\mathbb{E}\left[\|x_{t+1} - y_{t+1}\|_{P_{t+1}}^2\right].$$

It can be easily proven that

$$\mathbb{E}\left[\|g_t\|_{P_t}^{*2}\right] \leqslant 3\|\nabla f(x_t)\|_{P_t}^{*2} + 6L_t^2\|x_t - y_t\|_{P_t}^2.$$

We also need in the following lemma describing the properties of Hutchinson diagonal approximation

Lemma 11 (Jahani et al. [10]) *For L-Lipschitz smooth function f, it holds that*

1. $\left|\left[z_t \circ \nabla^2 f(x_{t+1}) z_t\right]_i\right| \leqslant \Gamma \leqslant \sqrt{n}L$.
2. $\exists \delta \leqslant 2(1-\beta)\Gamma$ *such that* $\forall t : \|P_{t+1} - P_t\|_\infty \leqslant \delta$.

Then, we can bound last term of V_{t+1} as follows

$$\mathbb{E}\left[\|x_{t+1} - y_{t+1}\|_{P_{t+1}}^2\right] \leqslant p\eta_t^2\mathbb{E}\left[\|g_t\|_{P_t}^{*2}\right] + (1-p)(1+\eta_t b_t)\|x_t - y_t\|_{P_t}^2$$
$$+ (1-p)\frac{\eta_t}{b_t}\|\nabla f(x_t)\|_{P_t}^{*2} + 2\Gamma\mathbb{E}\left[(1-\beta_{t+1})\|x_{t+1} - y_{t+1}\|_2^2\right]$$

due to Fenchel–Young inequality $\langle \nabla f(x_t), x_t - y_t\rangle \leq \frac{1}{b_t}\|\nabla f(x_t)\|^{*^2} + b_t\|x_t - y_t\|^2$ for some sequence $b_t > 0$, $t = 1, 2, ...$ we specify later. Now,

$$\begin{aligned}
\mathbb{E}\left[V_{t+1}\right] &\leqslant f(x_t) - f(x_*) - \eta_t\left(1 - (1-p)\frac{a_{t+1}}{b_t}\right)\|\nabla f(x_t)\|^{*^2}_{P_t} + a_{t+1}(1-p)(1+\eta_t b_t)\|x_t - y_t\|^2_{P_t} \\
&\quad + \eta_t^2\left(\frac{L_t}{2} + a_{t+1}\right)\left(3\|\nabla f(x_t)\|^{*^2}_{P_t} + 6L_t^2\|x_t - y_t\|^2_{P_t}\right) + 2a_{t+1}\Gamma\mathbb{E}\left[(1-\beta_{t+1})\|x_{t+1} - y_{t+1}\|_2^2\right] \\
&\leqslant f(x_t) - f(x_*) - \eta_t\left(1 - (1-p)\frac{a_{t+1}}{b_t} - 3a_{t+1}\eta_t - 3L_t\eta_t\right)\|\nabla f(x_t)\|^{*^2}_{P_t} \\
&\quad + a_{t+1}\left((1-p)(1+\eta_t b_t) + 3\eta^2\left(\frac{L_t}{a_{t+1}} + 2\right)L_t^2\right)\|x_t - y_t\|^2_{P_t} \\
&\quad + 2a_{t+1}\Gamma\mathbb{E}\left[(1-\beta_{t+1})\|x_{t+1} - y_{t+1}\|_2^2\right].
\end{aligned}$$

Finally, we determine the step size η_t satisfying

$$\begin{cases}
1 - (1-p)\dfrac{a_{t+1}}{b_t} - 3a_{t+1}\eta_t - 3L_t\eta_t \geqslant \dfrac{1}{4}, \\
(1-p)(1+\eta_t b_t) + 3\eta_t^2\left(\dfrac{L_t}{a_{t+1}} + 2\right)L_t^2 \leqslant \dfrac{a_t}{a_{t+1}}.
\end{cases}$$

We can set $a_{t+1} = L_t, b_t = p/\eta_t, \eta_t = c_t/L_t$ (this is only a comfortable option, but it is not unique; one can try to find an optimal one) for every $t = 1, 2, ...$ and variable sequence c_t, and after substitution we solve it with respect to c_t to obtain $\eta_t \leqslant \min\left\{\frac{\alpha p}{3}, \frac{3}{4}\frac{p}{5p+1}\right\}\frac{1}{L_t}$ for some $\alpha > 0$ as a sufficient condition[1]. Since the system of inequalities above holds now, we have

$$\mathbb{E}\left[V_{t+1}\right] \leqslant V_t - \frac{\eta_t}{4}\|\nabla f(x_t)\|^{*^2}_{P_t} + 2L_{t-1}\Gamma(1-\beta_{t+1})\|x_{t+1} - y_{t+1}\|_2^2,$$

that is summed up for $t = 1, ..., T$ to

$$\mathbb{E}\left[\|\nabla f(\overline{x}_T)\|^{*^2}_{P_t}\right] \leqslant \frac{1}{\sum_{t=1}^T \eta_t}\left[V_0 - V_* + 2\Gamma\sum_{t=2}^{T+1} L_{t-1}\mathbb{E}\left[(1-\beta_t)\|x_t - y_t\|_2^2\right]\right],$$

where $\overline{x}_T$ is such that $\overline{x}_T = x_t$ with probability $\eta_t / \sum_{k=1}^T \eta_k$ or, in the best-iteration manner,

$$\min_{t=1,...,T}\|\nabla f(x_t)\|^{*^2}_{P_t} \leqslant \frac{1}{\sum_{t=1}^T \eta_t}\left[V_0 - V_* + 2\Gamma\sum_{t=2}^{T+1} L_t\mathbb{E}\left[(1-\beta_t)\|x_t - y_t\|_2^2\right]\right].$$

It is time to remember that, in opposite to standard analysis, η_t depends on L_t in our case, so we have factor of the form $1/\left(\sum_{t=1}^T 1/L_t\right)$ in convergence rate. Writing it as harmonic average finishes the reasoning.

1. To be pedantic, $\eta_t \leqslant \min\left\{\frac{\sqrt{p^2 - 1 + L_{t-1}/L_t}}{3}, \frac{3}{4}\frac{p}{5p+1}\right\}\frac{1}{L_t}$, but it is necessary that $L_{t-1}/L_t \to 1$ on the one hand and L_t is close to 1 on the other hand, so we can find a proper $\alpha > 0$; note, that if we replace $a_{t+1} = L_t$ with $a_{t+1} = \max_{k=1,...,t} L_k$ we do not have any problems with $a_t/a_{t+1} \leqslant 1$, but need to set $\eta \propto 1/L$ that we do not want to do — this is why L_t is close to 1 is important condition.

B.7. Proof of Theorem 7

Firstly, let us rewrite the descent lemma in a way to get rid of cubic term, as follows

$$\mathbb{E}\left[f(x_{t+1})\right] \leqslant f(x_t) - \frac{\eta_t}{2}\|\nabla f(x_t)\|^{*2}_{P_t} + \frac{\eta_t}{2}(1-p)\left(1 + \frac{(1-\beta_{t+1})\beta_t}{1/\kappa_t + \beta_t\beta_{t+1}}\right)\|g_t - \nabla f(x_t)\|^{*2}_{P_t} + \frac{\eta_t}{2}\left[\left(1 + \frac{M'}{3}\sqrt{\frac{1-\beta_t\chi_t}{1+\sigma}}\|g_t\|^*_{P_t}\right)\left(\frac{\eta_t(1+\sigma)}{1-\beta_t\chi_t}\right)^2 + \frac{\eta_t(1+\sigma)}{1-\beta_t\chi_t} - 1\right]\|g_t\|^{*2}_{P_t}.$$

Denoting $\dfrac{\eta_t(1+\sigma)}{1-\beta_t\chi_t}$ as x and $\dfrac{M'}{3}\sqrt{\dfrac{1-\beta_t\chi_t}{1+\sigma}}$ as A, we get the $\dfrac{\eta_t}{2}\left((1 + A\|g_t\|^*_{P_t})x^2 + x - 1\right)$ factor for the term $\|g_t\|^{*2}_{P_t}$, which is less than zero and can be neglected if

$$0 < x \leqslant \frac{1}{\Phi + B(1-\beta_t\chi_t)^{1/4}\sqrt{\|g_t\|^*_{P_t}}} = \frac{2}{1+\sqrt{5}+2\sqrt{A\|g_t\|^*_{P_t}}} \leqslant \frac{-1+\sqrt{5+4A\|g_t\|^*_{P_t}}}{2(1+A\|g_t\|^*_{P_t})}, \tag{4}$$

where B denotes the $\sqrt{\dfrac{M'}{6}}(1+\sigma)^{-1/4} \leqslant \sqrt{\dfrac{M'}{6}}$ and $\Phi = \dfrac{1+\sqrt{5}}{2}$ is a golden ratio. This leads to the first upper bound on a step size:

$$\eta_t \leqslant \frac{1-\beta_t\chi_t}{(1+\sigma)\left(\Phi + B\sqrt{\|g_t\|^*_{P_t}}\right)} \leqslant \frac{1-\beta_t\chi_t}{(1+\sigma)\left(\Phi + B(1-\beta_t\chi_t)^{1/4}\sqrt{\|g_t\|^*_{P_t}}\right)}.$$

Secondly, to establish the decrease of variance term, we consider Lyapunov function $V_t := f(x_t) - f(x_*) + \|g_t - \nabla f(x_t)\|^{*2}_{P_t}$. Then, we need to upper bound step size once again to obtain

$$\frac{\eta_t}{2}(1-p)\left(1 + \frac{(1-\beta_{t+1})\beta_t}{1/\kappa_t + \beta_t\beta_{t+1}}\right) \leqslant 1.$$

Further, we need to distance β_{t+1} from zero. It would be to rough to lower bound all β with some fixed value, so we add the relation limiting the decreasing of β instead: $\beta_{t+1} \geqslant \beta_t/2$. This relation forces algorithm to be conservative and do not change preconditioner too much on a later iterations, that seems to be natural, because for a wide class of functions (self-concordant, for example) Hessian does not change much if step is small enough which holds for small gradients. So, upper bound on step size in our case looks like

$$\eta_t \leqslant \frac{1/(1+\kappa_t) + \beta_t}{1-p} \leqslant \frac{2/\kappa_t + \beta_t^2}{(1-p)\,(1/\kappa_t + \beta_t)} \leqslant \frac{2}{(1-p)\left(1 + \dfrac{(1-\beta_{t+1})\beta_t}{1/\kappa_t + \beta_t\beta_{t+1}}\right)}$$

Thus, if $\eta_t \leqslant \min\left\{\dfrac{1-\beta_t\chi_t}{(1+\sigma)\left(\Phi + \sqrt{M'/6\cdot\|g_t\|^*_{P_t}}\right)}, \dfrac{1/(1+\kappa_t)+\beta_t}{1-p}\right\}$, we have (now without accumulating errors!)

$$\mathbb{E}\left[V_{t+1}\right] \leqslant V_t - \frac{\eta_t}{4}\|\nabla f(x_t)\|^{*2}_{P_t} \implies \min_{t=1,\ldots,T}\|\nabla f(x_t)\|^{*2}_{P_t} = O\left(\frac{f(x_0) - f(x_*)}{\sum_{t=1}^T \eta_t}\right).$$

Appendix C. Supplementary numerical experiments

Firstly, we compare the performance of `Scaled L-SVRG` and ordinary `L-SVRG` on the problems with different smoothness characteristics A. For each value of A, we determine the best values of $\beta_t \equiv \beta$ and $\eta_t \equiv \eta$ for `Scaled L-SVRG`, and $\eta_t \equiv \eta$ for ordinary `L-SVRG` by logarithmically spaced grid search: $\eta \in \{2^{-2}, ..., 2^{-10}\}$, $\beta \in \{1 - 2^{-5}, ..., 1 - 2^{-10}\}$, while the precision achieved by tuned algorithm is estimated on average of 3 runs with random sequences of batches.

Figure 8 shows the results of comparison of `Scaled L-SVRG` and ordinary `L-SVRG` for $A \in \{0.1, 5, 10, 50\}$. Horizontal axis measures the number of iterations (stochastic gradient evaluations), vertical axis measures the objective function value $f(x_t)$, which is more practically interesting, or squared norm of the gradient $\|\nabla f(x_t)\|^2$, which is main for the theory in non-convex case. Convergence curves show the average value of quantity measured for 3 runs with random sequences of batches and are equipped with transparent shades of the size of standard deviation of the measurements. One can see that `Scaled L-SVRG` converge significantly faster than `L-SVRG` in all the cases. `Scaled L-SVRG` allows one to choose bigger step size even if its value is the same for all the iterations. Such a significant superiority of `Scaled L-SVRG` in the case of $A = 0.1$ might seem to be unexpected, because scaling with $A < 1$ leads to decreasing of Lipschitz constant and increasing of effective step size $\propto 1/L$ whilst scaling introduced by `Scaled L-SVRG` seeks to eliminate this effect. Nevertheless, scaling in algorithm turns out to be efficient through component-wise adaptivity — we encourage this effect by scaling features with random factors a parametrized by A.

Next experiment is devoted to the choice of step size $\eta_t \equiv \eta$ for different values of A ($A \in \{0.1, 5, 10, 50\}$). For each A, we set $\beta_t \equiv \beta$ to the best value determined for the `Scaled L-SVRG` in the previous experiment and consider $\eta \in \{2^{-4}, ..., 2^{-10}\}$. We also do not equip corresponding convergence curves with standard deviation shades in this experiment: it is not so significant here, and for most of runs one can estimate the scale of variance with the unaided eye.

Figure 9 shows the difference in convergence rate of `Scaled L-SVRG` in dependence on choice of step size. With the increasing of A (and hence L) convergence curves are pressed against the horizontal axis. Its natural, because effective step size is $\propto 1/L$, so efficiency of particular step size η, getting closer to effective step size, is improving as well, if η is small enough. Starting from $A = 50$, big step sizes, getting closer to the bound on a step size $\propto 1/L$ guaranteeing the compensation of variance, become inefficient.

Further, we focus on the behaviour of `Scaled L-SVRG` algorithm in dependence on the choice of $\beta_t \equiv \beta$, for varying A. We consider the case of constant β to validate the results obtained in "One-step effects" section.

Figure 10 summarizes the results of `Scaled L-SVRG` runs with $\beta \in \{0.95, 0.95 + \frac{1-0.95}{20}, ..., 1\}$. Horizontal axis measures value of β, vertical axis measures objective function value $f(x_T)$ (or squared norm of the gradient $\|\nabla f(x_T)\|^2$) after $T = 300$ iterations of the algorithm. Curves show the average value of quantity measures in 5 runs with random sequences of batches and are equipped with shades of the size of standard deviation of measurements. It can be seen that dependence of achieved precision on β changes with increasing of smoothness characteristic A: minimum of the corresponding function is getting closer to $\beta = 1$, its values on the left from minimum are growing and its growth rate near $\beta = 1$ is significantly increasing (which is especially noticeable for $A = 50$). This relationship between β and L (through A) reflects the trade-off between variance compensation and scaling gradients Lipschitz constant. Variance affects the convergence if β is small (this fact also leads to divergence for too small β's); increasing of L leads to the increasing of δ_t^+, which increases

the accumulating error term in (2), so, values for the small β's grow. To explain the behaviour near $\beta = 1$, it is reasonable to go to $A = 0.1$ case, where variance error terms are insignificant. Values begin to grow rapidly starting from $\beta \approx 0.97$ and stop on some fixed value at $\beta = 1$. This behaviour is described in (1), where we have shown the $O(1/(1-\beta))$ growth of gradients Lipschitz constant. The boundedness at $\beta = 1$ can be explained by the proper choice of P_0, such that $\delta_t^- \neq 1$, even if $\beta = 1$. Thus, the main outline of our theory is successfully confirmed on the experiment.

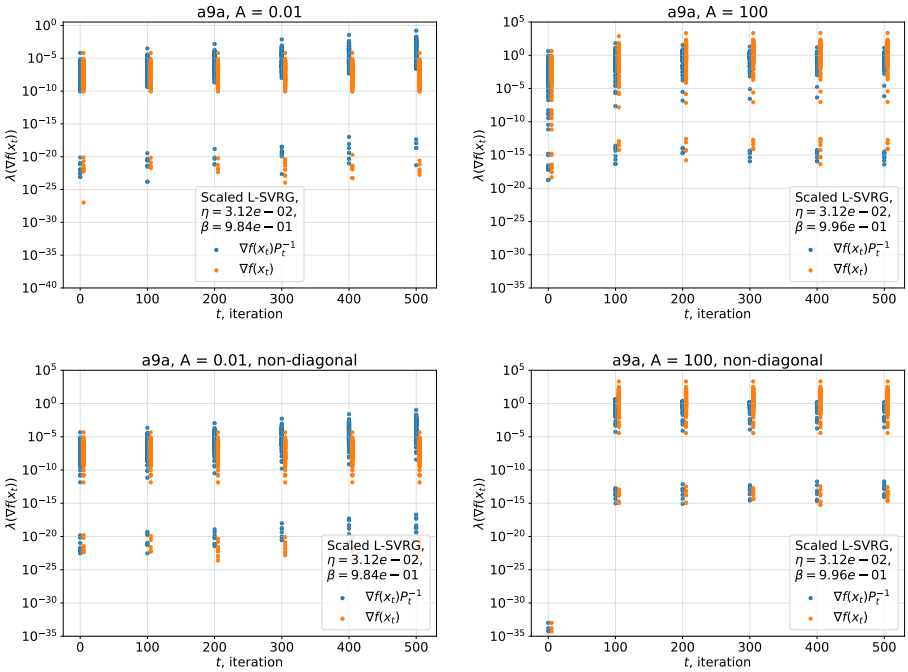

Figure 4: Dependence of spectrum of Hessian and $\nabla^2 f(x_t)P_t^{-1}$ (characterized by Δ_t) on number of iterations, for diagonal and non-diagonal preconditioning.

Next part of the experiments is about the patterns in preconditioners and related inexactnesses changing, and corresponding effect on the convergence of `Scaled L-SVRG` algorithm. Firstly, we consider the dynamic of the spectrum of Hessian and scaled Hessian, that is $\nabla^2 f(x_t)P_t^{-1}$, with increasing number of iterations. The main inexactness Δ_t upper bounds the largest eigenvalue of the scaled Hessian (minus one), and we do not present curve for Δ_t, because one can estimate it up to the order with the unaided eye. This is a first time we consider non-diagonal preconditioning in our experiments; such an update is defined by $d_t = |\nabla^2 f_{B_t}(x_t)|_\epsilon$, with the same B_t and ϵ, which now requires singular value decomposition of $\nabla^2 f_{B_t}(x_t)$ at the every iteration. Strictly speaking, our theory is not well-suited to this case, but this practical consideration will give us an additional information about behaviour of the algorithm when the smallest and other eigenvalues are scaled in a proper way.

Figure 4 presents the dynamic of Hessian and scaled Hessian spectrum in two scenarios: diagonal and non-diagonal, for $A \in \{0.01, 100\}$, which is needed to represent both $A < 1$ and $A > 1$ cases. What we see is that the largest eigenvalue of scaled Hessian converges to 1, which means its increasing in comparison to the largest eigenvalue of Hessian in the case of $A < 1$ and its decreasing — in the case of $A > 1$. For the problem we consider, all the eigenvalues of the (scaled) Hessian form two clouds of points of the plot, and whilst the upper cloud is shifted so that the largest

eigenvalue tends to 1, relative position of lower cloud depends on the update type. If we use a diagonal update, lower cloud is shifted in the same direction as the upper one (all the eigenvalues increase or decrease at the same time in the case of $A < 1$ or $A > 1$, correspondingly), which leads to the moving of the smallest eigenvalue of the scaled Hessian away from 1. It worth to note that diagonal update do not pay enough attention to small eigenvalues that can be seen also from Figure 11. Conversely, if we use non-diagonal update, clouds can be shifted in opposite directions such that both the largest and the smallest eigenvalues of the scaled Hessian converge to 1.

The following Figure 5 show the results of the experiments, similar to ones described for Figure 3 before, that is, present the dependencies of optimal β_* on smoothness characteristic A and number of iterations T, but for non-diagonal preconditioner. In comparison with analogous dependencies for diagonal updates, optimal β_* for non-diagonal updates are significantly less sensitive to the change of A and T. In particular, dependence of β_* on A in non-diagonal case is closer to linear (because it is closer to exponential in logarithmic scale for A), so the previous remark on sensitivity of β_* to the change of $A \ll 1$ ceases to be relevant. Similarly, the growth of β_* with increasing t is significantly slower than in diagonal case and is also less monotonic. Taking into account the range of β values on both figures, one can say that $\beta_* = (0.965 \pm 0.005)$ independently on smoothness of the problem and the number of iterations. Thus, our hypothesis is that the faster the smallest eigenvalue (together with the largest one) of the scaled Hessian tends to 1 with increasing number of iterations, the less dependent optimal β_* is on the smoothness and number of iterations, which means in the extreme case that optimal β_* is determined by some affine-invariant characteristic of the function. Note that such a β_* can be greater than β_* obtained for diagonal updates (cf. Figure 3).

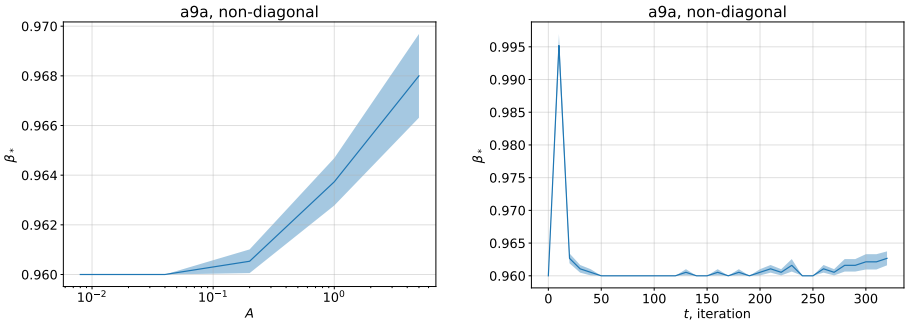

Figure 5: Dependencies of optimal $\beta_t \equiv \beta$ for non-diagonal preconditioning.

The last group of experiments is related to the version of `Scaled L-SVRG` with the step sizes chosen with line-search. We use Brent algorithm which searches for $\eta_t \in [0, 1]$ with minimal value of the $f(x_t - \eta_t P_t^{-1} g_t)$ by only function evaluations. This could be non-practical, if calculation of all the objective function's terms is computationally expensive, but in the case when the most expensive operation is evaluation of the gradient it is acceptable. Besides, our interest to `Scaled L-SVRG` with line-search is more theoretical — namely, by this modification we would like to reach the advantage of scaling introduced by averaging of the smoothness constants. Indeed, in previous experiments the step size was fixed, such that Lipschitz constant of the gradient was included in convergence rate as a true constant — scaled, but not better than by fixed preconditioner. Here, on the contrary, algorithm exploits scaling as much as possible. So, we compare the performance of algorithms with or without line-search to assess this advantage, and show the dependence of performance on β.

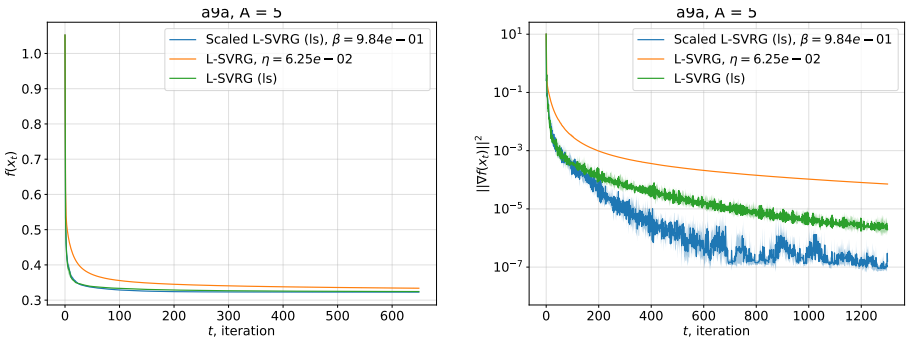

Figure 6: Convergence curves of `L-SVRG`, `L-SVRG` with line-search and `Scaled L-SVRG` with line-search with optimal choice of $\beta_t \equiv \beta$ and $\eta_t \in [0, 1]$.

On the Figure 6, we compare `L-SVRG`, `L-SVRG` with line-search (with the aim of fair comparison) and `Scaled L-SVRG` with line-search. Firstly, precision obtained by the algorithms with line-search is much better. At the same time, performance of `L-SVRG` with line-search and `Scaled L-SVRG` with line-search is almost the same until 200 iterations, and advantage of scaling comes clear with increasing of number of iterations and improving the preconditioner (see Figure 4). It is natural: the average of smoothness constants decreases with adapting of the preconditioner.

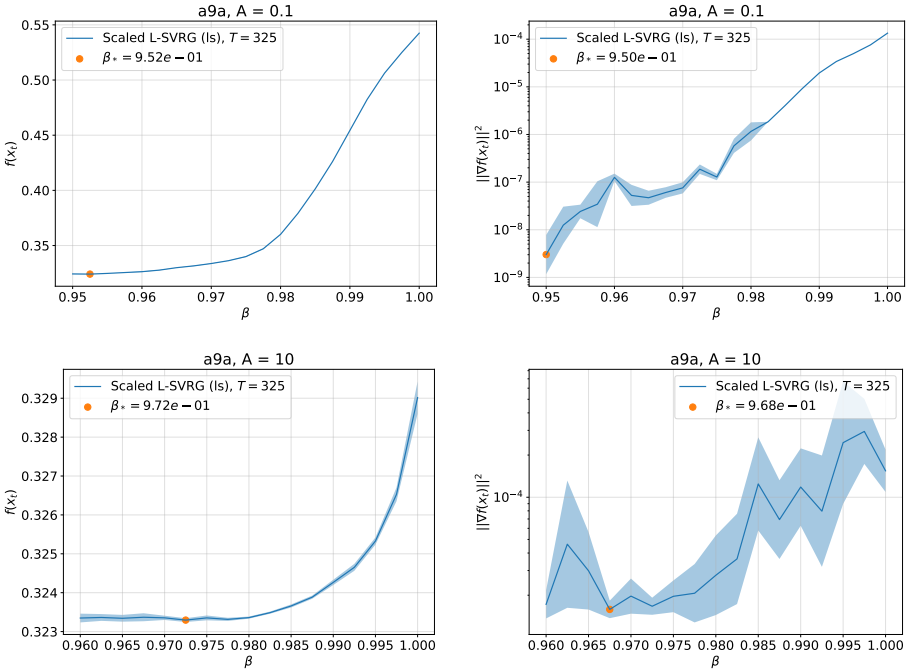

Figure 7: Dependence of achieved precision on $\beta_t \equiv \beta$ with line-search.

Then, we reproduce the comparison of the `Scaled L-SVRG` operation in dependence on β. The results are shown on the Figure 7 (cf. Figure 10). Summarizing the differences, small β values became acceptable even for big values of A, so that preconditioner can adapt faster without sacrificing convergence rate. This is a little unexpected, because the convergence slowdown for small β values is explained primarily by the variance introduced by changing preconditioner, and the use of line-search does not relieve us of this factor. For now, we cannot explain this effect with certainty.

Appendix D. Omitted figures

D.1. Experiments for LibSVM `a9a` dataset

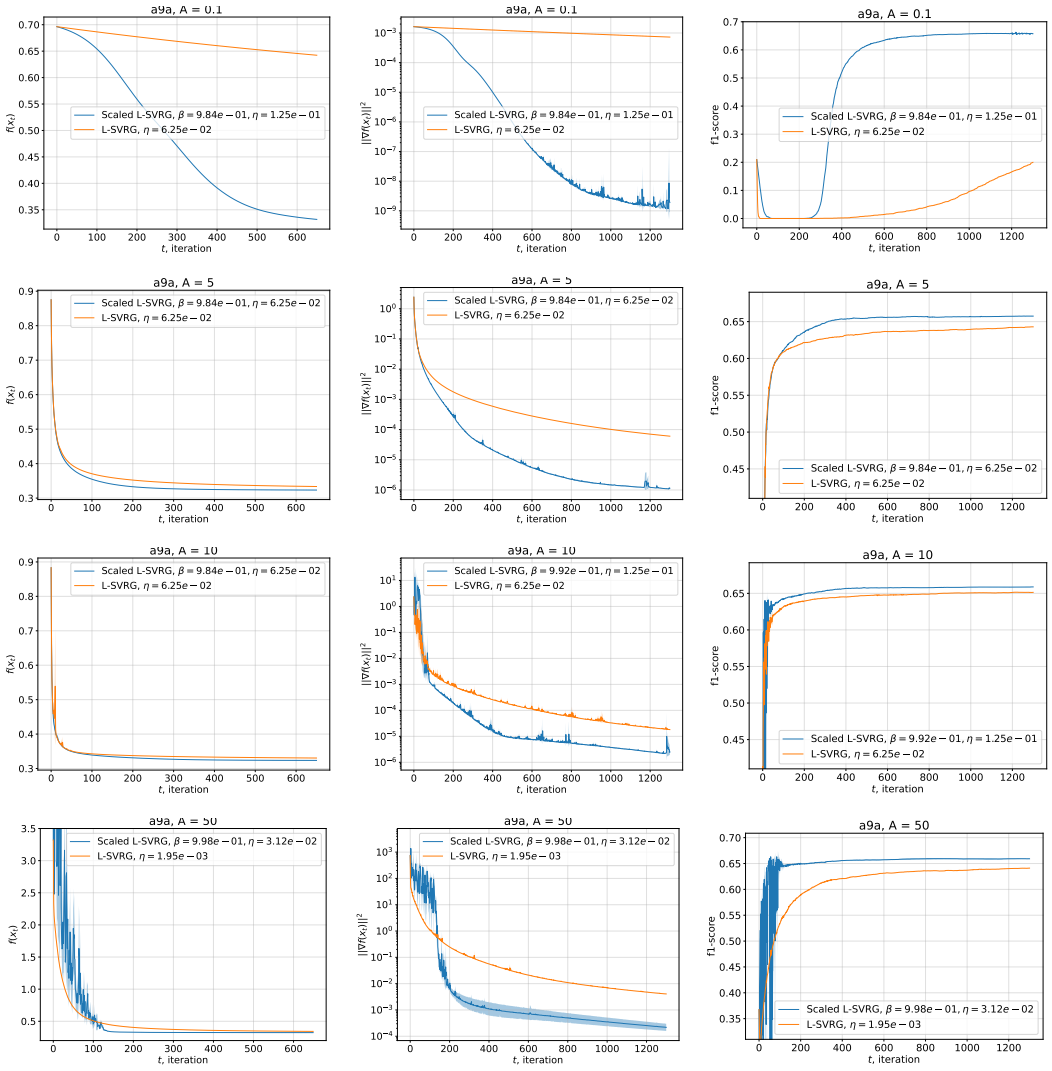

Figure 8: Convergence curves of `L-SVRG` and `Scaled L-SVRG` with optimal choice of $\beta_t \equiv \beta$ and $\eta_t \equiv \eta$.

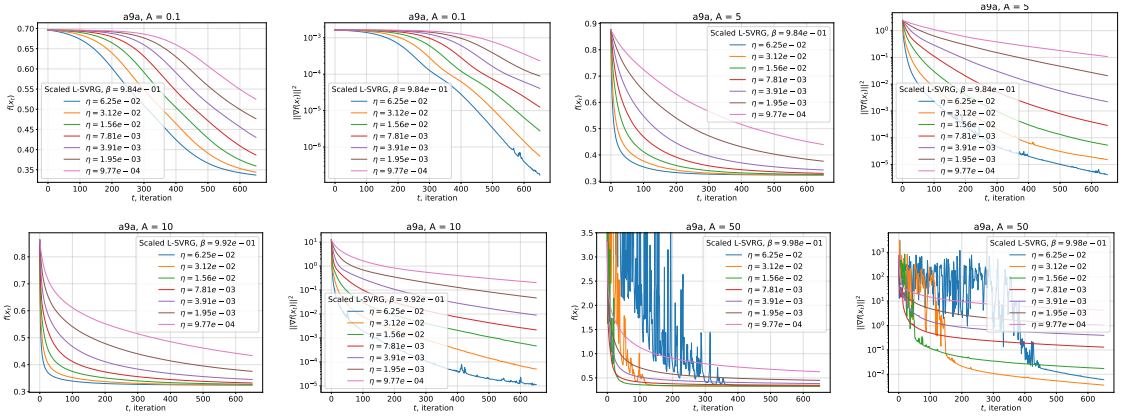

Figure 9: Convergence curves of `Scaled L-SVRG` with different step sizes on logistic regression problems with different Lipschitz constants.

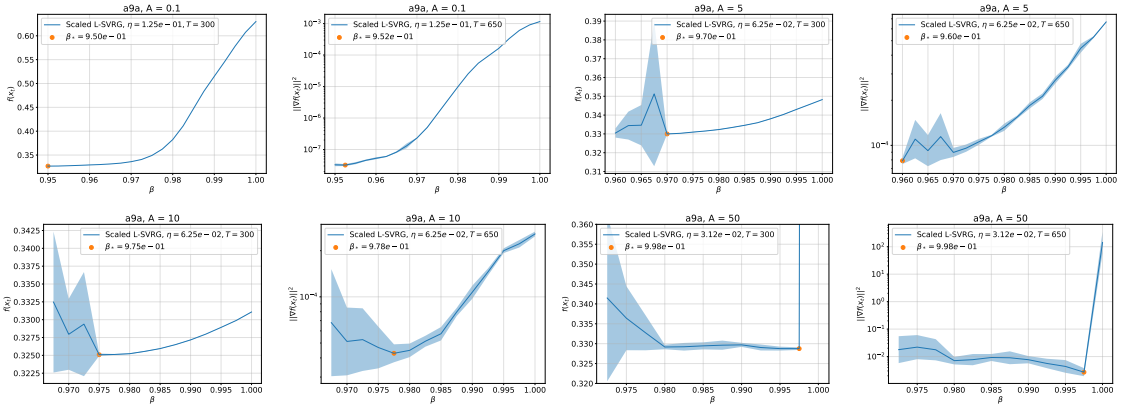

Figure 10: Dependence of achieved precision on $\beta_t \equiv \beta$. (Note: on the lower left plot, function value at $\beta = 1$ is too big, so we cropped the picture, that is why the curve is almost vertical there)

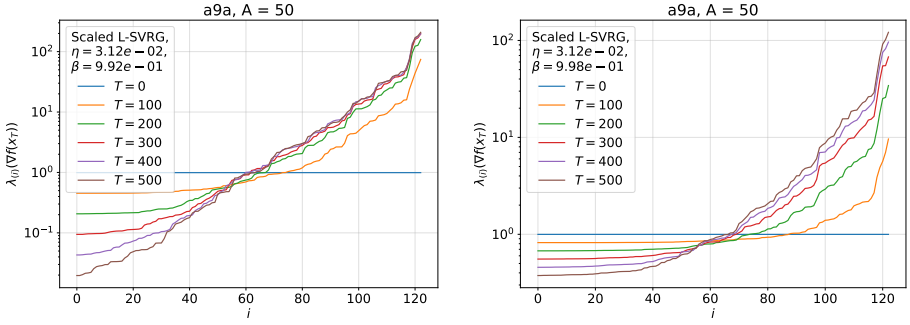

Figure 11: Dependence of spectrum of Hessian approximation on number of iterations.

D.2. Experiments for LibSVM `covtype-binary-scaled` dataset

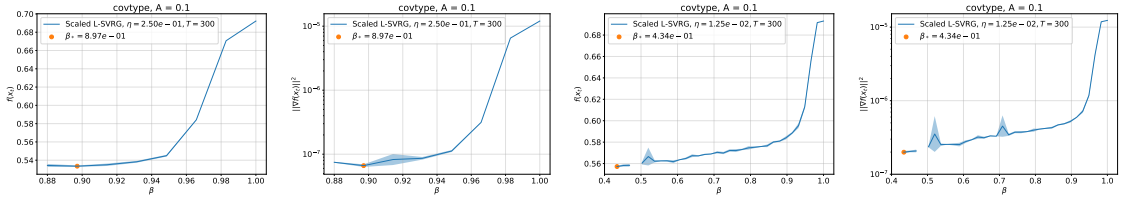

Figure 12: Dependence of achieved precision on $\beta_t \equiv \beta$, $A = 0.1$. (Note: gaps in curves mean the divergence of the algorithm in at least one of 3 runs)

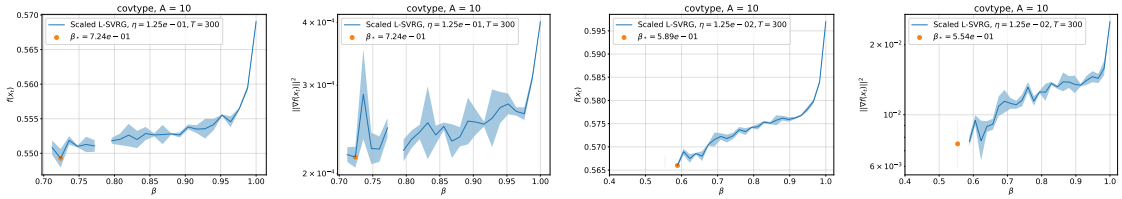

Figure 13: Dependence of achieved precision on $\beta_t \equiv \beta$, $A = 10$. (Note: gaps in curves mean the divergence of the algorithm in at least one of 3 runs)

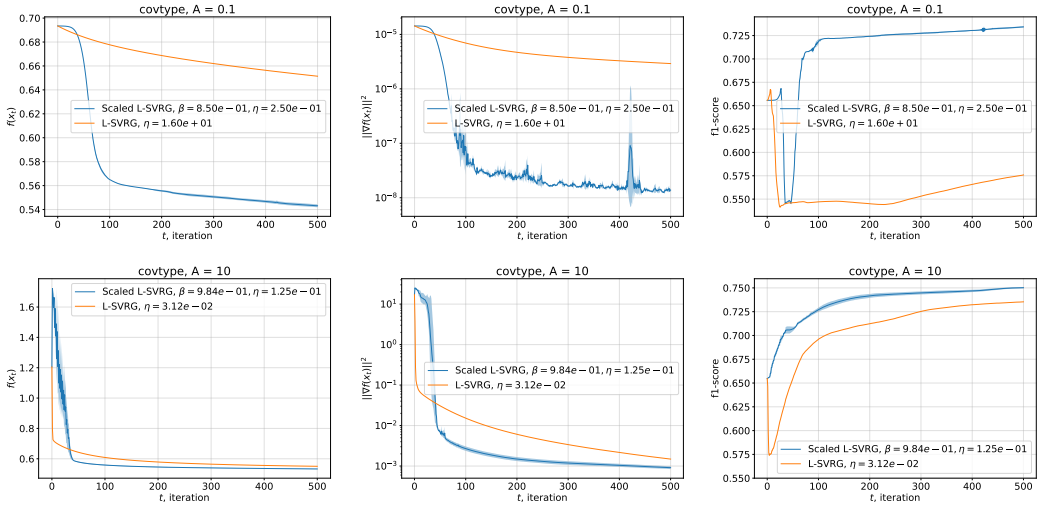

Figure 14: Convergence curves of `L-SVRG` and `Scaled L-SVRG` with optimal choice of $\beta_t \equiv \beta$ and $\eta_t \equiv \eta$.