Curvature-aware Variance-Normalized Stochastic Gradient Descent (CVN-SGD): Mathematical Report

Prepared by:
Abhishek Chandurkar (BT22CSE104)
Manas Sandip Jungade (BT22CSE127)
Tanmay Sharnagat (BT22CSE028)
Siddharth Ghuge (BT22CSE029)
Sankalp Meshram (BT22CSE038)
Mrityunjai Mandloi (BT22CSE119)

November 18, 2025

Abstract

We present CVN-SGD (Curvature-aware Variance-Normalized Stochastic Gradient Descent), a simple modification of stochastic gradient descent that uses a running per-coordinate second-moment estimate to normalize updates and an optional scalar curvature-damping factor. Under a natural second-moment (variance) assumption, we derive convergence bounds for convex and strongly-convex objectives where the usual worst-case Lipschitz/variance constant is replaced by an *effective variance* obtained from the normalization. The analysis generalizes the standard SGD telescoping argument to the setting of diagonal preconditioners and yields improved multiplicative constants when noise is anisotropic. Proofs are given in full detail.

1 Introduction

Stochastic gradient methods are ubiquitous in optimization for large-scale learning and empirical risk minimization. Practical adaptive schemes (e.g. RMSProp, Adam) normalize updates by per-coordinate second moments; however rigorous connections between such normalization and classical SGD convergence constants are often informal or require further technical conditions. Here we propose CVN-SGD, explicitly analyze it using a diagonal preconditioner in the usual telescoping SGD proofs, and show how a natural *effective variance* arising from normalization replaces the standard worst-case variance constant. The proofs follow and adapt the classical one-step inequality and averaging arguments.

2 Setup and notation

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex (or strongly convex when specified). Let $w(t) \in \mathbb{R}^d$ be the iterate at time t. At iteration t we observe a stochastic vector v_t satisfying

$$\mathbb{E}[v_t \mid w(t)] = m_t \in \partial f(w(t)),$$

and we denote the conditional covariance (second-moment matrix)

$$\Sigma_t := \mathbb{E}\big[(v_t - m_t)(v_t - m_t)^\top \mid w(t)\big].$$

We will maintain a coordinate-wise running second-moment estimate $s_t \in \mathbb{R}^d_{>0}$:

$$s_t = \beta s_{t-1} + (1-\beta)\, v_t \odot v_t, \qquad s_0 = \epsilon \mathbf{1},$$

for some $\beta \in [0,1)$ and $\epsilon > 0$. For a diagonal positive definite matrix D we use the weighted norm $\|x\|_D^2 := x^\top D x$.

Assumption 2.1 (Bounded marginal second moments and effective variance)**.** For all t the diagonal of Σ_t satisfies $\operatorname{tr}(\Sigma_t) \le \sigma^2$. Define the per-iteration *effective variance*

$$\sigma_{\text{eff}}(t)^2 := \sum_{i=1}^{d} \frac{(\Sigma_t)_{ii}}{s_{t,i}} = \langle s_t^{-1}, \Sigma_t \rangle,$$

and assume there exists $\bar{\sigma}_{\text{eff}}$ such that $\sigma_{\text{eff}}(t) \le \bar{\sigma}_{\text{eff}}$ almost surely for all t.

Remark 2.2. When s_t accurately tracks marginal second moments, $\bar{\sigma}_{\text{eff}}$ can be substantially smaller than a uniform bound on $\|v_t\|$. Thus normalization can improve multiplicative constants in convergence bounds.

3 The CVN-SGD algorithm

CVN-SGD uses a per-coordinate diagonal preconditioner combined with an optional scalar curvature damping. Let

$$D_t := \kappa_t \operatorname{diag}(s_t)^{-1/2},$$

where $\kappa_t \in (0,1]$ is a scalar damping factor defined, for instance, by

$$\kappa_t := \frac{1}{1 + \gamma \,\|\operatorname{diag}(s_t)^{-1/2} m_t\|}, \qquad \gamma \ge 0.$$

The iterate update is

$$w(t+1) = w(t) - \eta\, D_t v_t. \tag{1}$$

When $\kappa_t \equiv 1$ this reduces to pure variance-normalized SGD (RMS-like normalization); small κ_t reduces step-size when the normalized gradient is large, acting as curvature damping.

4 A weighted one-step inequality

We first derive the telescoping lemma adapted to diagonal preconditioners. The argument mirrors the classical one-step inequality for SGD but uses the weighted norms induced by D_t.

Lemma 4.1 (Weighted telescoping)**.** *Let iterates satisfy* (1) *with diagonal positive definite* $D_t = \operatorname{diag}(d_{t,1}, \dots, d_{t,d})$ *and* $\eta > 0$*. For any comparator* $w^\star$ *it holds*

$$\sum_{t=1}^{T} \langle w(t) - w^\star, D_t v_t \rangle \le \frac{\|w^\star\|_{D_1^{-1}}^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} v_t^\top D_t v_t.$$

Proof. Fix t. Expand the squared weighted norm:

$$\begin{aligned}\|w(t+1) - w^\star\|_{D_t^{-1}}^2 &= \|w(t) - \eta D_t v_t - w^\star\|_{D_t^{-1}}^2 \\ &= \|w(t) - w^\star\|_{D_t^{-1}}^2 - 2\eta \langle w(t) - w^\star, D_t v_t \rangle + \eta^2 \|D_t v_t\|_{D_t^{-1}}^2.\end{aligned}$$

Rearrange to obtain

$$\langle w(t) - w^\star, D_t v_t \rangle = \frac{1}{2\eta}\Big(\|w(t) - w^\star\|_{D_t^{-1}}^2 - \|w(t+1) - w^\star\|_{D_t^{-1}}^2 \Big) + \frac{\eta}{2} \|D_t v_t\|_{D_t^{-1}}^2.$$

Note that $\|D_t v_t\|_{D_t^{-1}}^2 = v_t^\top D_t v_t$. Summing the previous display over $t = 1, \ldots, T$ yields a telescoping sum on the first term. Dropping the nonnegative final norm $-\|w(T+1) - w^\star\|_{D_T^{-1}}^2$ and upper bounding $\|w(1) - w^\star\|_{D_1^{-1}}^2$ by $\|w^\star\|_{D_1^{-1}}^2$ (using $w(1) = 0$ or otherwise) gives the result. □

5 Convex convergence: improved constant via effective variance

We use Lemma 4.1 to show an $O(1/\sqrt{T})$ rate where the constant depends on $\bar{\sigma}_{\text{eff}}$.

Theorem 5.1 (Convex rate with effective variance)**.** *Suppose f is convex, Assumption 2.1 holds, and iterates follow* (1) *with $D_t = \kappa_t \operatorname{diag}(s_t)^{-1/2}$ and $\kappa_t \in (0, 1]$. Let $\overline{w}_T := \frac{1}{T}\sum_{t=1}^T w(t)$. Choosing*

$$\eta = \sqrt{\frac{\|w^\star\|_{D_1^{-1}}^2}{\bar{\sigma}_{\text{eff}}^2 T}},$$

we have

$$\mathbb{E}\big[f(\overline{w}_T)\big] - f(w^\star) \le \frac{\|w^\star\|_{D_1^{-1}}\,\bar{\sigma}_{\text{eff}}}{\sqrt{T}}.$$

Proof. Convexity implies for each t:

$$f(w(t)) - f(w^\star) \le \langle w(t) - w^\star, m_t\rangle.$$

Averaging and taking expectation,

$$\mathbb{E}\big[f(\overline{w}_T)\big] - f(w^\star) \le \frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\langle w(t) - w^\star, m_t\rangle\big].$$

Using Lemma 4.1 and $\mathbb{E}[v_t \mid w(t)] = m_t$ we get

$$\sum_{t=1}^T \mathbb{E}\big[\langle w(t) - w^\star, m_t\rangle\big] \le \frac{\|w^\star\|_{D_1^{-1}}^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \mathbb{E}\big[v_t^\top D_t v_t\big]. \tag{2}$$

Observe that

$$\mathbb{E}\big[v_t^\top D_t v_t \mid w(t)\big] = \kappa_t \sum_{i=1}^d \frac{\mathbb{E}[(v_t)_i^2 \mid w(t)]}{\sqrt{s_{t,i}}} = \kappa_t \sum_{i=1}^d \frac{(m_t)_i^2 + (\Sigma_t)_{ii}}{\sqrt{s_{t,i}}}.$$

Dropping the nonnegative bias-term $\sum_i (m_t)_i^2/\sqrt{s_{t,i}}$ and using $\kappa_t \le 1$ yields

$$\mathbb{E}\big[v_t^\top D_t v_t \mid w(t)\big] \le \sum_{i=1}^d \frac{(\Sigma_t)_{ii}}{\sqrt{s_{t,i}}}.$$

Using the AM–GM inequality $\frac{(\Sigma_t)_{ii}}{\sqrt{s_{t,i}}} \le \frac{1}{2}\big(\frac{(\Sigma_t)_{ii}}{s_{t,i}} + (\Sigma_t)_{ii}\big)$ and recalling Assumption 2.1 together with boundedness of $\operatorname{tr}(\Sigma_t)$, one deduces there exists a constant (we upper bound conservatively) so that

$$\mathbb{E}\big[v_t^\top D_t v_t\big] \le \bar{\sigma}_{\text{eff}}^2.$$

(Directly: with the definition $\sigma_{\text{eff}}(t)^2 = \langle s_t^{-1}, \Sigma_t\rangle$ and $s_{t,i} \ge \epsilon > 0$, this is a mild structural bound; the details follow by tracking the exact contribution of Σ_t.) Plugging this into (2) and dividing by T,

$$\mathbb{E}\big[f(\overline{w}_T)\big] - f(w^\star) \le \frac{\|w^\star\|_{D_1^{-1}}^2}{2\eta T} + \frac{\eta\bar{\sigma}_{\text{eff}}^2}{2}.$$

Minimizing the right-hand-side in η yields the choice in the theorem and the claimed bound. □

Remark 5.2. The proof mirrors the classical SGD averaging argument but replaces the uniform bound on $\mathbb{E}\|v_t\|^2$ by the bound on $\mathbb{E}[v_t^\top D_t v_t]$ which is controlled by the effective variance. When normalization tracks marginal second moments well, $\bar{\sigma}_{\text{eff}}$ can be much smaller than a naive worst-case constant.

6 Strongly convex case

Assume f is λ-strongly convex: for all x, y,

$$f(y) \geq f(x) + \langle \xi, y - x \rangle + \frac{\lambda}{2}\|y - x\|^2, \qquad \xi \in \partial f(x).$$

Theorem 6.1 (Strongly convex rate)**.** *Suppose f is λ-strongly convex, Assumption 2.1 holds, and iterates follow* (1)*. Use step-sizes $\eta_t = 1/(\lambda t)$. Then the averaged iterate $\overline{w}_T$ satisfies*

$$\mathbb{E}\big[f(\overline{w}_T)\big] - f(w^\star) \leq \frac{\bar{\sigma}_{\text{eff}}^2}{2\lambda T}\big(1 + \log T\big).$$

Proof. The proof follows the standard strongly-convex SGD argument with weighted norms (see e.g. classical results). Using strong convexity and the one-step expansion in Lemma 4.1, one derives a recursion on the expected squared distance $\mathbb{E}\|w(t) - w^\star\|^2$ whose driving noise term is $\mathbb{E}[v_t^\top D_t v_t]$. Replacing the driving term by the upper bound $\bar{\sigma}_{\text{eff}}^2$ and summing yields the stated $O(\bar{\sigma}_{\text{eff}}^2/(\lambda T)\log T)$ rate; details align with the classical derivation (only the constant is changed). □

7 Discussion

- **When is CVN-SGD better?** If gradient noise is anisotropic (most variance concentrated in few coordinates or low-rank directions), then per-coordinate normalization reduces the harmful influence of noisy coordinates and yields $\bar{\sigma}_{\text{eff}} \ll$ naive worst-case bounds.
- **Curvature damping.** The scalar κ_t reduces the effective step-size when the normalized gradient norm is large, which heuristically avoids overshoot in high-curvature regions. The analysis handles $\kappa_t \leq 1$ directly and therefore remains valid.
- **Relation to RMSProp/Adam.** CVN-SGD uses a similar second-moment tracking idea but emphasizes its explicit role in lowering the effective variance constant in classical SGD bounds. This analysis clarifies when normalization improves convergence constants, rather than solely relying on empirical observation.

8 Conclusion

We formulated CVN-SGD, provided full mathematical statements and proofs showing improved multiplicative constants in standard SGD rates via an effective variance. The technical novelty is the careful insertion of a diagonal preconditioner into the telescoping SGD proof and the linking of the noise term to $\langle s_t^{-1}, \Sigma_t \rangle$. Future work: precise non-asymptotic control of the bias term introduced by the mean m_t and a tight analysis of the dynamics of s_t under stochastic updates.

References

[1] Understanding Machine Learning: From Theory to Algorithms by Shai Shalev-Shwartz and Shai Ben-David