On-Average Stability of Multipass Preconditioned SGD and Effective Dimension

Simon Vary[1] Tyler Farghly[1] Ilja Kuzborskij[2] Patrick Rebeschini[1]

[1]University of Oxford [2]Google DeepMind

COLT 2026

When does preconditioning help or hurt generalisation?

Slides: simonvary.com/colt26.pdf

Setup: Preconditioned SGD

Minimise the population risk $f(x) = \mathbb{E}_{z\sim Q}[\ell(x,z)]$ from n i.i.d. samples $S = \{z_1, \dots, z_n\}$ minimizing $f_S(x) = \frac{1}{n}\sum_i \ell(x, z_i)$ with **multipass** preconditioned SGD:

$$x_{t+1} = x_t - \eta_t\, P\, \nabla\ell(x_t, z_{i_t}), \qquad i_t \sim \mathrm{Unif}\{1, \dots, n\}.$$

Three sources of curvature govern the problem:

- H **Loss curvature** - Hessian of the loss $\nabla^2\ell \approx H, \quad \nabla^2\ell(x,z) \preceq \beta H$
- Σ **Gradient noise** - covariance $\mathrm{Cov}_z\left(\nabla\ell(x,z)\right) \preceq \Sigma$
- P **Preconditioner** - fixed by the algorithm

Question

How does the excess risk $\mathbb{E}[\, f(x_t) - \min f\,]$ depend on the interplay of H, Σ, and P?

When noise and curvature disagree

In the well-specified case, i.e.,
$\ell(x^*, z) \coloneqq -\log p_Q(z|x^*)$,

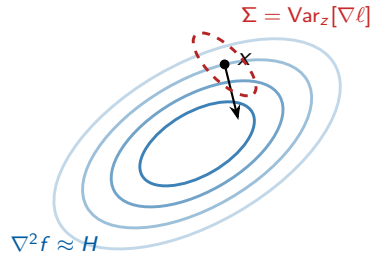
$$\Sigma \;=\; H \;=\; \text{Fisher information.}$$

Then *natural grad.* $P \approx H^{-1}$ is optimal [Amari '98].

But models almost always misspecified: $\Sigma \neq H$.

Optimisers have *different* targets:

- **Adam / K-FAC** $\rightarrow$ whiten the *second-moments* ($P = \mathbb{E}[\nabla\ell\nabla\ell^\top]^{-1} \approx \Sigma^{-1}$)
- **AdaHessian / SketchySGD** $\rightarrow$ invert the *curvature* ($P \approx H^{-1}$)



Curvature (H) $\neq$ noise (Σ).

Whitening the second-moments can destabilise high-curvature directions.

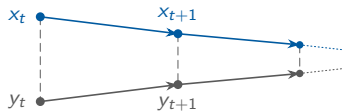
Main result: excess risk of PSGD

Theorem

For contractive updates, step size $\eta_t \sim 1/t^\gamma$, where $\gamma \in (0,1]$ and $n \gtrsim \kappa_\ell\, \kappa(PH)$:

$$\mathbb{E}[f(x_t) - \min f] \;\lesssim\; \underbrace{\beta\, \frac{\operatorname{tr}(PHP\,\Sigma)}{t^\gamma}}_{\text{optimisation} \sim 1/t^\gamma} \;+\; \underbrace{\operatorname{tr}(P\,\Sigma)\Big(\frac{1}{\sqrt{n\,t^\gamma}} + \frac{1}{n}\Big)}_{\text{generalisation} \sim 1/n},$$

Contractive updates



Optimal preconditioner

$P = H^{-1}$, $\gamma = 1$ $\Rightarrow$ *both* terms collapse to

$$\operatorname{tr}\big(H^{-1}\Sigma\big)\,\left(\frac{1}{t} + \frac{1}{n}\right).$$

Recovers the optimal TIC rate.

The same geometry controls both

- Optimisation variance $\rightarrow$ $\operatorname{tr}(PHP\,\Sigma)$
- Statistical / stability term $\rightarrow$ $\operatorname{tr}(P\,\Sigma)$

Second-order information is not only optimization — it is robustness to sampling noise.

Beyond contractivity: At convergence

Theorem

For f_S satisfying μ-quadratic growth (non-convex), and $n \gtrsim \beta/\mu$:

$$\mathbb{E}[f(x_t) - \min f] \;\leq\; \frac{2\beta}{\mu}\,\mathbb{E}[f_S(x_t) - \min f_S] \;+\; \frac{16}{\mu}\,\frac{\operatorname{tr}\big(H^{-1}\Sigma\big)}{n}.$$

At convergence the bound is the optimal effective dimension — independent of P.

$P = H^{-1}$ is the sweet spot: fastest optimisation *and* smallest effective dimension.

The right complexity measure: effective dimension

Classical statistics replaces the ambient dimension d by the effective dimension

$$\boxed{\operatorname{tr}\left(H^{-1}\Sigma\right)}$$

also known as the Takeuchi Information Criterion (TIC).

It is the *optimal* statistical rate $\operatorname{tr}\left(H^{-1}\Sigma\right)/n$ for:

- exact ridge-regression minimisers [Bach '24],
- averaged SGD [Neu & Rosasco '18],
- asymptotic stochastic approximation [Polyak & Juditsky '92] (Cramér–Rao).

This work

We bring $\operatorname{tr}\left(H^{-1}\Sigma\right)$ into the *non-asymptotic* analysis of **multipass PSGD** and show how the choice of P can *break* it.

On-average algorithmic stability

Standard approach is worst-case: Say x'_t is trained on $S' = S \cup \{z'\} \setminus \{z_i\}$

$$\mathbb{E}_S\big[f(x_t) - f_S(x_t)\big] = \underbrace{\mathbb{E}_{S,z'}\big[\ell(x'_t, z_i) - \ell(x_t, z_i)\big]}_{\text{on-average stability}} \leq L\,\mathbb{E}_{S,z'}\|x_t - x'_t\|,$$

but, this usual uniform bound is a worst-case sample.

Instead, to reveal the geometry:

- use β-smoothness (*not* Lipschitzness)
- on-average (not uniform) stability $\Rightarrow$ the bound *sees* the noise geometry Σ.

$$\mathbb{E}_S\big[f(x_t) - f_S(x_t)\big] \lesssim \underbrace{\mathbb{E}\big[\|\nabla\ell(x_t, z')\|_*^2\big]^{1/2}}_{\text{gradient/noise geometry } \Sigma} \cdot \underbrace{\mathbb{E}\big[\|x_t - x'_t\|^2\big]^{1/2}}_{\text{parameter stability } \textbf{squared}}.$$

Stability for *multipass* SGD

Aim: Bound $\mathbb{E}\|x_t - x_t'\|_M^2$ where $\|x\|_M := \sqrt{x^\top M x}$ for $M \succ 0$.

Assumption: r-contractivity of PSGD in $\|\cdot\|_M$-norm, β-smoothness.

Lemma

Provided $n \gtrsim \kappa_\ell\, \kappa(PH)$, *the correlated term is benign and unrolls to*

$$\mathbb{E}_{S,z'}\left[\|x_t - x_t'\|_M^2\right] \lesssim \frac{\operatorname{tr}(PMP\Sigma)}{n^2} + \eta\,\frac{\operatorname{tr}(PMP\Sigma)}{n}.$$

Proof idea: the perturbed iterate has never seen z_i

The two runs differ only when index i is drawn (prob. $1/n$):

$$\mathbb{E}\|x_{t+1} - x'_{t+1}\|_M^2 \lesssim (1 - \eta\, r)\, \mathbb{E}\|x_t - x'_t\|_M^2 + \frac{\eta^2}{n}\, \mathbb{E}\big\|\nabla \ell(x_t, z_i) - \nabla f(x_t)\big\|_{PMP}^2.$$

Problem: after first pass $x_t \not\perp\!\!\!\perp z_i$.

Key observation: $x'_t \perp\!\!\!\perp z_i$ and $x'_t \perp\!\!\!\perp z_i$ $\Rightarrow$ Swap iterates, pay with smoothness:

$$\nabla \ell(x_t, z_i) - \nabla f(x_t) = \underbrace{\nabla \ell(x'_t, z_i) - \nabla f(x'_t)}_{\text{fresh variance} \le \operatorname{tr}(PMP\Sigma)} + \underbrace{\big[\nabla \ell(x_t, z_i) - \nabla \ell(x'_t, z_i)\big] - \big[\nabla f(x_t) - \nabla f(x'_t)\big]}_{\text{smoothness: } \lesssim \beta \|x_t - x'_t\|}$$

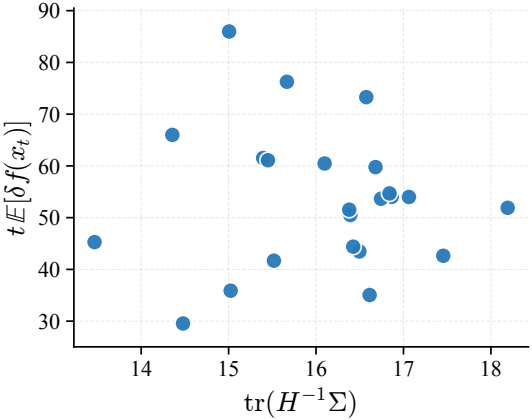
Contraction balances the correlation cost:

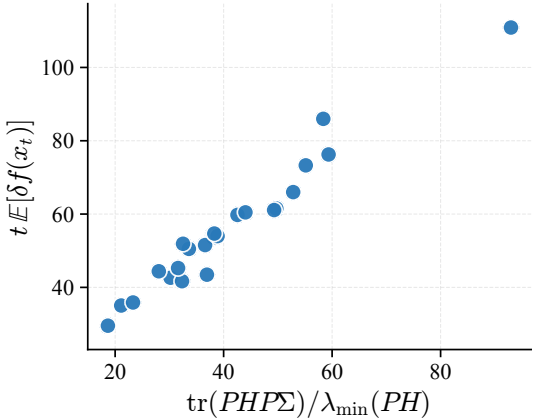
$$\mathbb{E}\|x_{t+1} - x'_{t+1}\|_M^2 \lesssim \left(1 - \eta r + \frac{\eta\beta}{n}\right) \mathbb{E}\|x_t - x'_t\|_M^2 + \frac{\eta^2}{n} \operatorname{tr}(PMP\,\Sigma),$$

So for $n \gtrsim \beta/r$ the contraction absorbs it $\Longrightarrow$ same as if all samples were fresh.

Experiment: it's the interplay, not H, Σ alone

Noisy quadratics, 24 random instances (H, Σ, P) (non-commuting), $d = 10$. Plot scaled last-iterate risk $t\,\mathbb{E}[\delta f(x_t)]$ against two predictors.





Non-asymptotic behaviour is governed by (P, H, Σ) jointly — not by H, Σ alone.

Takeaways

1. **New tool:** on-average stability analysis for multipass SGD handles correlated iterates, no Lipschitz assumption.
2. **Geometry of generalisation:** excess risk of PSGD is controlled by an effective dimension jointly shaped by curvature H, noise Σ, and preconditioner P.
3. $P = H^{-1}$ **is optimal** for *both* optimisation *and* generalisation — recovering the optimal rate $\operatorname{tr}\big(H^{-1}\Sigma\big)\big(1/t + 1/n\big)$, *even under misspecification*.
4. **A bad** P **hurts:** matching instance-dependent lower bounds show the constant can blow up by $\kappa\big(PH\big)$, even with decaying steps.

Thank you! `simon.vary@stats.ox.ac.uk`

References I

- Amari (1998). *Natural gradient works efficiently in learning*. Neural Computation.
- Hardt, Recht, Singer (2016). *Train faster, generalize better*. ICML.
- Kuzborskij, Lampert (2018). *Data-dependent stability of SGD*. ICML.
- Thomas et al. (2020). *On the interplay between noise and curvature*. AISTATS.
- Bach (2024). *Learning Theory from First Principles*. MIT Press.
- Neu, Rosasco (2018). *Iterate averaging as regularization*. COLT.
- Polyak, Juditsky (1992). *Acceleration of stochastic approximation by averaging*. SIAM J. Control Optim.
- Karimi, Nutini, Schmidt (2016). *Linear convergence under the PL condition*. ECML.

Matching lower bounds: a bad P really does hurt

Minimax (any estimator). The statistical floor is exactly the effective dimension:

$$\inf_{\hat{x}} \sup_{P} \mathbb{E}[\delta f(\hat{x})] \gtrsim \frac{\operatorname{tr}\big(H^{-1}\Sigma\big)}{n\,\alpha}.$$

Instance-dependent (PSGD, decaying step $\eta_t \sim 1/t$). The last iterate obeys

$$\mathbb{E}[\delta f(x_t)] \gtrsim \frac{\operatorname{tr}(PHP\,\Sigma)}{\alpha\,\lambda_{\max}(PH)\,\lambda_{\min}(PH)} \cdot \frac{1}{t} + \frac{\operatorname{tr}\big(H^{-1}\Sigma\big)}{\alpha} \cdot \frac{1}{n}.$$

The price of misalignment

- $P = H^{-1}$: matches the upper bound (up to κ_ℓ).
- Ill-conditioned P ($\kappa(P) = 1/\varepsilon$): risk $\gtrsim \operatorname{tr}(H\Sigma)/(\varepsilon\, t)$ — constant blows up *even with decaying steps*.
- Plain SGD ($P = I$): a factor $\kappa(H)$ worse than optimal.

Minimax is too coarse here — the *instance* (the choice of P) decides performance.