What Makes a Good Preconditioner for Data Science?

Mitchell T. Scott

Department of Mathematics, Emory University

41st Colorado Conference On Iterative and Multigrid Methods

CCIMM 2026

August 6, 2026

Collaborators and Acknowledgments

Yousef Saad,
Minnesota

Tianshi Xu,
~~Emory~~ Kentucky

Yuanzhe Xi,
Emory

This work was partially supported by NSF DMS-2208412 and NSF DMS-2513118.

Table of Contents

Stochastic gradient descent is the workhorse of ML training optimizers.

Example

I want to minimize some loss, but the dataset is so large, the gradient computation takes forever!
What do I do?

Stochastic gradient descent is the workhorse of ML training optimizers.

Example

I want to minimize some loss, but the dataset is so large, the gradient computation takes forever!
What do I do?

Definition (Preconditioned Stochastic Gradient Descent)

To minimize the stochastic optimization problem, we can write the resulting algorithm as

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \mathbf{M}^{-1} g(\mathbf{w}_k, \boldsymbol{\xi}_k), \tag{1}$$

where $g(\mathbf{w}_k, \boldsymbol{\xi}_k) = \nabla_w F_k(\mathbf{w})$ is the stochastic gradient, α_k is the learning rate, $\boldsymbol{\xi}_k$ is an i.i.d. sample drawn at iteration k, and $\mathbf{M}$ is the preconditioner.

EMORY UNIVERSITY

Common preconditioners for SGD can consider isotropy and curvature.

Example (Adam, Kingma and Ba, 2017 [1])

Define $\mathbf{s}_k$ is an exponential moving average of squared gradients up until iteration k, then

$$\mathbf{M}_{\text{Adam}} = \operatorname{diag}\left(\sqrt{\mathbf{s}_k} + \boldsymbol{\varepsilon}\right),$$

Common preconditioners for SGD can consider isotropy and curvature.

Example (Adam, Kingma and Ba, 2017 [1])

Define $\mathbf{s}_k$ is an exponential moving average of squared gradients up until iteration k, then

$$\mathbf{M}_{\text{Adam}} = \operatorname{diag}\left(\sqrt{\mathbf{s}_k} + \boldsymbol{\varepsilon}\right),$$

Example (Full or mini-batch Hessian)

The Hessian, or mini-batch approximation to the Hessian, matrix of the loss function,

$$\mathbf{H}_{\mathcal{B}}(\mathbf{w}) = \nabla^2 \mathcal{L}_{\mathcal{B}}(\mathbf{w}).$$

Common preconditioners for SGD can consider isotropy and curvature.

Example (Adam, Kingma and Ba, 2017 [1])

Define $\mathbf{s}_k$ is an exponential moving average of squared gradients up until iteration k, then

$$\mathbf{M}_{\text{Adam}} = \operatorname{diag}\left(\sqrt{\mathbf{s}_k} + \boldsymbol{\varepsilon}\right),$$

Example (Full or mini-batch Hessian)

The Hessian, or mini-batch approximation to the Hessian, matrix of the loss function,

$$\mathbf{H}_{\mathcal{B}}(\mathbf{w}) = \nabla^2 \mathcal{L}_{\mathcal{B}}(\mathbf{w}).$$

Example (Fisher Information Matrix (FIM))

Using the FIM as a preconditioner corresponds to natural gradient descent [Amari, 1998].

$$\mathbf{F} = \mathbb{E}\left[\frac{\mathrm{d}\log p(y|x;\mathbf{w})}{\mathrm{d}\mathbf{w}} \frac{\mathrm{d}\log p(y|x;\mathbf{w})}{\mathrm{d}\mathbf{w}}^{\top}\right] := \mathbb{E}[\mathcal{D}\mathbf{w}\mathcal{D}\mathbf{w}^{\top}] = \mathbb{C}\mathrm{ov}(\mathcal{D}\mathbf{w}, \mathcal{D}\mathbf{w})$$

EMORY UNIVERSITY

Table of Contents

Classical Convergence Results for SGD

Theorem (Convergence of SGD, Robbins and Monro, 1951 [3])

Assume that a loss function F where $F_* := F(\mathbf{w}^*)$, has L-Lipschitz gradients and $\mathbb{E}\left[\|\nabla F(\mathbf{w})\|^2\right] \leq \sigma^2$, then SGD with fixed step size $(\alpha_k = \overline{\alpha})$ converges

$$\mathbb{E}\left[F(\mathbf{w}_k) - F_*\right] \leq \frac{F(\mathbf{w}_1) - F_*}{2\overline{\alpha}k}$$

Classical Convergence Results for SGD

Theorem (Convergence of SGD, Robbins and Monro, 1951 [3])

Assume that a loss function F where $F_* := F(\mathbf{w}^*)$, has L-Lipschitz gradients and $\mathbb{E}\left[\|\nabla F(\mathbf{w})\|^2\right] \leq \sigma^2$, then SGD with fixed step size ($\alpha_k = \overline{\alpha}$) converges

$$\mathbb{E}\left[F(\mathbf{w}_k) - F_*\right] \leq \frac{F(\mathbf{w}_1) - F_*}{2\overline{\alpha}k}$$

Theorem (Strongly Convex Objective, Fixed Stepsize (Thm 4.6, Bottou, Curtis, and Nocedal, 2018 [4]))

Adding in stronger bounds on the gradient's first and second moments, and c-strong convexity. Running SGD with a small fixed step size $\overline{\alpha}$ satisfy

$$\mathbb{E}[F(\mathbf{w}_k) - F_*] \leq (1 - \overline{\alpha}c\mu)^{k-1}\left(F(\mathbf{w}_1) - F_* - \frac{\overline{\alpha}LK}{2c\mu}\right) + \frac{\overline{\alpha}LK}{2c\mu}, \tag{2}$$

where μ, K are constants associated with the stochastic gradients and variance, respectively.

Two important quantities for data science preconditioning.

Remark (Two stage convergence)

This highlights two late-stage drivers: a linear contraction factor $(1 - \alpha c\mu)$ *and a stochastic error floor*

$$\frac{\alpha L K}{2c\mu} = \frac{\alpha}{2\mu}\,\kappa\, K,$$

where $\kappa := \frac{L}{c}$ *is the (Euclidean) condition number associated with curvature.*

Two important quantities for data science preconditioning.

Remark (Two stage convergence)

This highlights two late-stage drivers: a linear contraction factor $(1-\alpha c\mu)$ and a stochastic error floor

$$\frac{\alpha L K}{2c\mu} = \frac{\alpha}{2\mu}\kappa K,$$

where $\kappa := \frac{L}{c}$ is the (Euclidean) condition number associated with curvature.

Definition (Conditional Variance)

$$\mathbb{V}_{\boldsymbol{\xi}_k}\left[g(\mathbf{w}_k, \boldsymbol{\xi}_k), \|\cdot\|_{\mathbf{M}^{-1}}\right] := \mathbb{E}_{\boldsymbol{\xi}_k}\left[\|g(\mathbf{w}_k, \boldsymbol{\xi}_k)\|^2_{\mathbf{M}^{-1}}\right] - \left\|\mathbb{E}_{\boldsymbol{\xi}_k}[g(\mathbf{w}_k, \boldsymbol{\xi}_k)]\right\|^2_{\mathbf{M}^{-1}} = \operatorname{tr}\left(\mathbf{M}^{-1}\boldsymbol{\Sigma}(\mathbf{w})\right), \quad (3)$$

where $\boldsymbol{\Sigma}(\mathbf{w}) := \mathbb{C}\mathrm{ov}(g(\mathbf{w}, \boldsymbol{\xi}) \mid \mathbf{w})$. For late stage convergence

$$\mathbb{V}_{\boldsymbol{\xi}_k}\left[g(\mathbf{w}_k, \boldsymbol{\xi}_k), \|\cdot\|_{\mathbf{M}^{-1}}\right] \leq K + \mathcal{O}(\overline{\alpha} K),$$

EMORY UNIVERSITY

Preconditioning helps convex problems converge faster with a lower loss.

Theorem (S., *et al.*, '26a, [5])

Running (1) *with small* $\alpha_k = \overline{\alpha}$ *under preconditioned analogues of common assumptions, then, for all* $k \in \mathbb{N}$,

$$\mathbb{E}[F(\mathbf{w}_k) - F_*] \leq \frac{\overline{\alpha}\,\hat{L}\,K}{2\,\hat{c}\,\mu} + (1 - \overline{\alpha}\,\hat{c}\,\mu)^{k-1}\left(F(\mathbf{w}_1) - F_* - \frac{\overline{\alpha}\,\hat{L}\,K}{2\,\hat{c}\,\mu}\right) \xrightarrow{k\to\infty} \frac{\overline{\alpha}\,\hat{L}\,K}{2\,\hat{c}\,\mu}. \tag{4}$$

Preconditioning helps convex problems converge faster with a lower loss.

Theorem (S., *et al.*, '26a, [5])

Running (1) *with small* $\alpha_k = \overline{\alpha}$ *under preconditioned analogues of common assumptions, then, for all* $k \in \mathbb{N}$,

$$\mathbb{E}[F(\mathbf{w}_k) - F_*] \leq \frac{\overline{\alpha}\,\hat{L}\,K}{2\,\hat{c}\,\mu} + (1 - \overline{\alpha}\,\hat{c}\,\mu)^{k-1}\left(F(\mathbf{w}_1) - F_* - \frac{\overline{\alpha}\,\hat{L}\,K}{2\,\hat{c}\,\mu}\right) \xrightarrow{k\to\infty} \frac{\overline{\alpha}\,\hat{L}\,K}{2\,\hat{c}\,\mu}. \tag{4}$$

Theorem

Now suppose (1) *has a decaying learning rate, then, for all* $k \in \mathbb{N}$,

$$\mathbb{E}[F(\mathbf{w}_k) - F_*] \leq \frac{\nu}{\gamma + k}, \qquad \nu := \max\left\{\frac{\beta^2\hat{L}K}{2(\beta\hat{c}\mu - 1)}, (\gamma+1)\big(F(\mathbf{w}_1) - F_*\big)\right\}. \tag{5}$$

EMORY UNIVERSITY

Polyak-Łojasiewicz functions are structured but nonconvex.

Assumption (Local PŁ functions on $\mathcal{N}$)

There exists $\hat{\mu}_{\mathrm{PL}} > 0$ *such that, for all* $\mathbf{w} \in \mathcal{N}$: $\quad 2\hat{\mu}_{\mathrm{PL}}\big(F(\mathbf{w}) - F_*\big) \leq \|\nabla F(\mathbf{w})\|^2.$

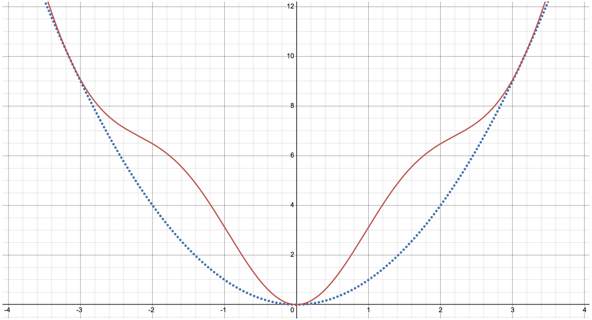


Figure 1: Comparing a convex function $F(x) = x^2$ with a $\mu = \frac{1}{40} -$ PŁ function $F(x) = x^2 + 3\sin^2(x)$.

Preconditioning helps in the nonconvex case, more realistic for SciML problems.

Theorem (S., *et al.*, '26a (Informal) [5])

Assume $\mathbf{w}_1$ *is in a quadratically growing local basin of radius* r, $\mathcal{N}_r$*. Let* $\tau := \inf\{k \geq 1 : \mathbf{w}_k \notin \mathcal{N}_r\}$.

Preconditioning helps in the nonconvex case, more realistic for SciML problems.

Theorem (S., *et al.*, '26a (Informal) [5])

Assume $\mathbf{w}_1$ *is in a quadratically growing local basin of radius* r, $\mathcal{N}_r$. *Let* $\tau := \inf\{k \geq 1 : \mathbf{w}_k \notin \mathcal{N}_r\}$.
For a fixed step size $\alpha_k = \overline{\alpha}$,

Preconditioning helps in the nonconvex case, more realistic for SciML problems.

Theorem (S., *et al.*, '26a (Informal) [5])

Assume $\mathbf{w}_1$ *is in a quadratically growing local basin of radius* r, $\mathcal{N}_r$. *Let* $\tau := \inf\{k \geq 1 : \mathbf{w}_k \notin \mathcal{N}_r\}$.
For a fixed step size $\alpha_k = \overline{\alpha}$,

- *we have linear convergence (depending on* $\uparrow \hat{\mu}_{\text{PL}}$*)*

Preconditioning helps in the nonconvex case, more realistic for SciML problems.

Theorem (S., *et al.*, '26a (Informal) [5])

Assume $\mathbf{w}_1$ *is in a quadratically growing local basin of radius* r, $\mathcal{N}_r$. *Let* $\tau := \inf\{k \geq 1 : \mathbf{w}_k \notin \mathcal{N}_r\}$.

For a fixed step size $\alpha_k = \overline{\alpha}$,

- *we have linear convergence (depending on* $\uparrow \hat{\mu}_{\mathrm{PL}}$*)*
- *and convergence up until a noise floor, which depends on a tradeoff between*
  - $\downarrow \hat{L}/\hat{\mu}_{\mathrm{PL}}$ *and*

Preconditioning helps in the nonconvex case, more realistic for SciML problems.

Theorem (S., *et al.*, '26a (Informal) [5])

Assume $\mathbf{w}_1$ *is in a quadratically growing local basin of radius* r, $\mathcal{N}_r$. *Let* $\tau := \inf\{k \geq 1 : \mathbf{w}_k \notin \mathcal{N}_r\}$.

For a fixed step size $\alpha_k = \overline{\alpha}$,

- *we have linear convergence (depending on* $\uparrow \hat{\mu}_{\mathrm{PL}}$*)*
- *and convergence up until a noise floor, which depends on a tradeoff between*
  - $\downarrow \hat{L}/\hat{\mu}_{\mathrm{PL}}$ *and*
  - $\downarrow K$.

Preconditioning helps in the nonconvex case, more realistic for SciML problems.

Theorem (S., *et al.*, '26a (Informal) [5])

Assume $\mathbf{w}_1$ *is in a quadratically growing local basin of radius* r, $\mathcal{N}_r$. *Let* $\tau := \inf\{k \geq 1 : \mathbf{w}_k \notin \mathcal{N}_r\}$.

For a fixed step size $\alpha_k = \overline{\alpha}$,

- *we have linear convergence (depending on* $\uparrow \hat{\mu}_{\mathrm{PL}}$*)*
- *and convergence up until a noise floor, which depends on a tradeoff between*
  - $\downarrow \hat{L}/\hat{\mu}_{\mathrm{PL}}$ *and*
  - $\downarrow K$.

For a harmonic step size $\alpha_k \propto 1/k$

Preconditioning helps in the nonconvex case, more realistic for SciML problems.

Theorem (S., *et al.*, '26a (Informal) [5])

Assume $\mathbf{w}_1$ *is in a quadratically growing local basin of radius* r, $\mathcal{N}_r$. *Let* $\tau := \inf\{k \geq 1 : \mathbf{w}_k \notin \mathcal{N}_r\}$.

For a fixed step size $\alpha_k = \overline{\alpha}$,

- *we have linear convergence (depending on* $\uparrow \hat{\mu}_{\mathrm{PL}}$*)*
- *and convergence up until a noise floor, which depends on a tradeoff between*
    - $\downarrow \hat{L}/\hat{\mu}_{\mathrm{PL}}$ *and*
    - $\downarrow K$.

For a harmonic step size $\alpha_k \propto 1/k$

- *we have* $\mathcal{O}(\nu/k)$ *convergence, where* ν *depends on* $\downarrow \hat{L}/\hat{\mu}_{\mathrm{PL}}$.

Preconditioning helps in the nonconvex case, more realistic for SciML problems.

Theorem (S., *et al.*, '26a (Informal) [5])

Assume $\mathbf{w}_1$ is in a quadratically growing local basin of radius r, $\mathcal{N}_r$. Let $\tau := \inf\{k \geq 1 : \mathbf{w}_k \notin \mathcal{N}_r\}$.

For a fixed step size $\alpha_k = \overline{\alpha}$,

- *we have linear convergence (depending on $\uparrow \hat{\mu}_{\mathrm{PL}}$)*
- *and convergence up until a noise floor, which depends on a tradeoff between*
  - $\downarrow \hat{L}/\hat{\mu}_{\mathrm{PL}}$ *and*
  - $\downarrow K$.

For a harmonic step size $\alpha_k \propto 1/k$

- *we have $\mathcal{O}(\nu/k)$ convergence, where ν depends on $\downarrow \hat{L}/\hat{\mu}_{\mathrm{PL}}$.*

We stay in the basin w.h.p. depending on $1 - \mathcal{O}\left(\hat{L}K\right) - \sum_k \delta_k$, where δ_k is the "escape probability".

Table of Contents

Diagnostic quadratic model allows for explicit control of preconditioned spectra.

Example (Quadratic model)

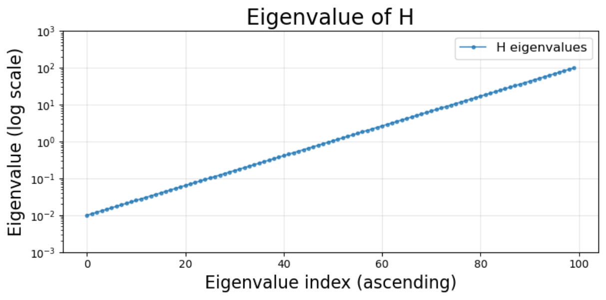
Let $\mathbf{w}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{w}_* = \mathbf{0}, F_* = 0$ for the quadratic model:

$$F(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}_*),$$

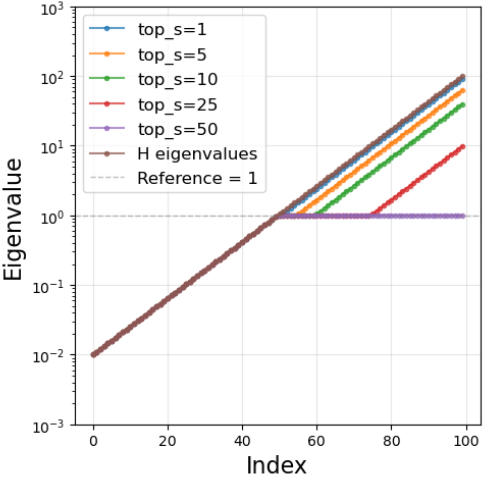
where $\mathbf{H} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$ has log-uniform spectrum from $[10^{-2}, 10^2]$, namely $\mathbf{H} \succ 0$.
We add stochastic noise, so

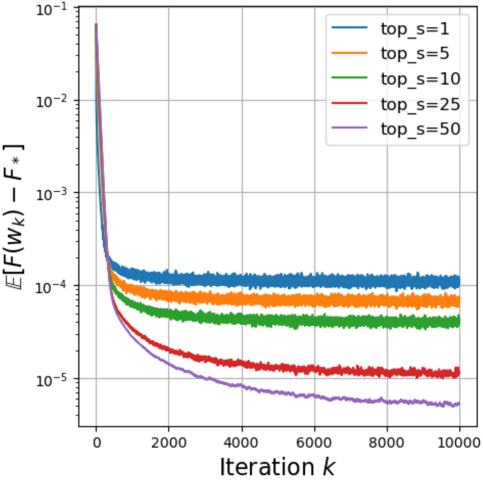
$$\nabla F(\mathbf{w}) = g(\mathbf{w}) + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

and run (preconditioned) SGD.

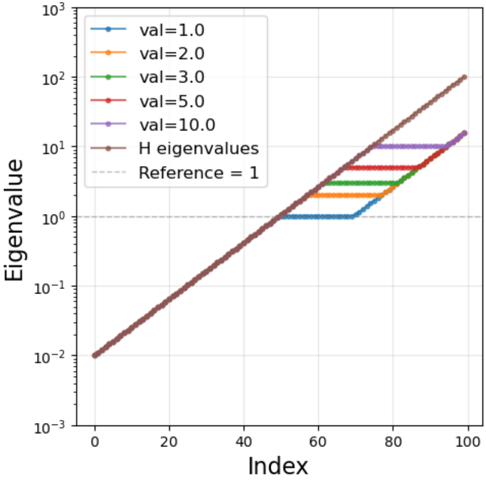


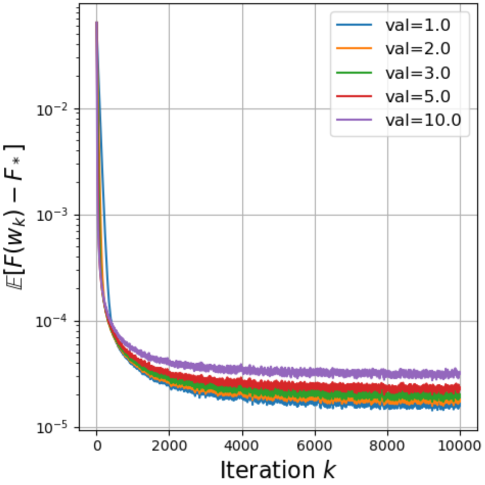
Deflation preconditioners lead to faster convergence and lower loss floor.





Generic deflation preconditioners demonstrate that a lower trace leads to a lower loss floor.

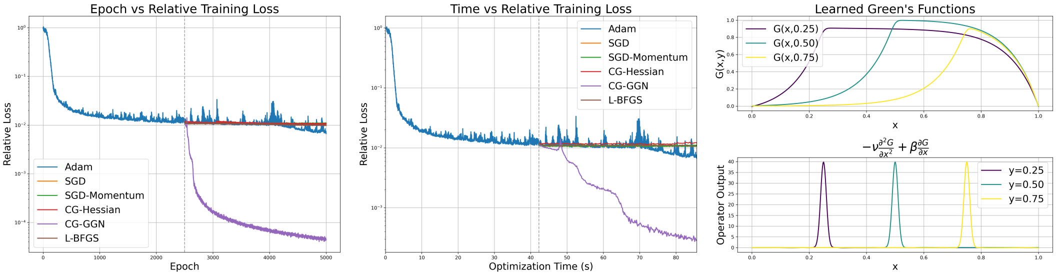




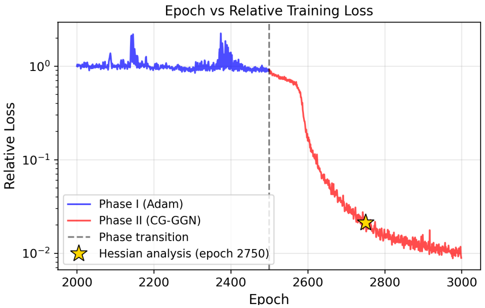
Convection dominated problems are hard, and Green's function learning is no different.

Example (Convection-dominated problem from Hao et al. 2024, [6])

Find $G(x,y)$ such that $\mathcal{L}G = \delta(x-y)$ where $\mathcal{L}u := -0.1u'' + 1.0u'$, $u(0) = u(1) = 0$



Preconditioning empirically clusters the eigenvalues for SciML's Hessian.



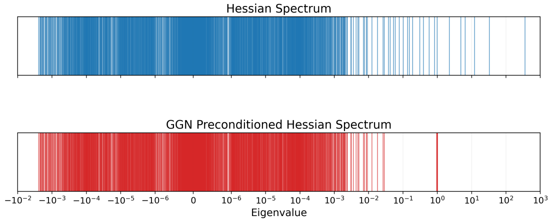


Table of Contents

Studying the Geometry of the (preconditioned) Loss Landscape

Example (Nonlinear Least Squares)

Given data points $(x_i, y_i), i = 1, \ldots, m \in \mathbb{R}^n$ let $\mathbf{r}_i(\mathbf{w}) := f(x_i, \mathbf{w}) - y_i$ be the loss of a nonlinear least squares problem. The Hessian is

$$\mathbf{H} := \mathbf{J}(\mathbf{w})\mathbf{J}(\mathbf{w})^\top + \text{small residuals}$$

where $\mathbf{J}$ is the Jacobian of $\mathbf{r}$,

Studying the Geometry of the (preconditioned) Loss Landscape

Example (Nonlinear Least Squares)

Given data points $(x_i, y_i), i = 1, \dots, m \in \mathbb{R}^n$ let $\mathbf{r}_i(\mathbf{w}) := f(x_i, \mathbf{w}) - y_i$ be the loss of a nonlinear least squares problem. The Hessian is

$$\mathbf{H} := \mathbf{J}(\mathbf{w})\mathbf{J}(\mathbf{w})^\top + \text{small residuals}$$

where $\mathbf{J}$ is the Jacobian of $\mathbf{r}$,

Remark

As $n \to \infty$, we (Pennington and Bahri, 2017 [7]) could view the Hessian as

$$\mathbf{H} := \mathbf{H}_0 + \mathbf{H}_1 = \frac{1}{T}\mathbf{X}\mathbf{X}^\top + \frac{1}{2}\left(\mathbf{W} + \mathbf{W}^\top\right), \quad \textit{where } \mathbf{X}, \mathbf{W} \sim \mathcal{N}(0,1)$$

How do we precondition this Loss Surface?

Let $\mathbf{X} \in \mathbb{R}^{d\times T}$ and $\mathbf{Y} \in \mathbb{R}^{d\times S}$ be free, standard Gaussian matrices, and define

$$\mathbf{H}_0 = \frac{1}{T}\mathbf{X}\mathbf{X}^\top, \qquad \mathbf{M} = \frac{1}{S}\mathbf{Y}\mathbf{Y}^\top, \qquad d/T \to q \in (0,1), \qquad d/S \to \psi \in (0,1).$$

How do we precondition this Loss Surface?

Let $\mathbf{X} \in \mathbb{R}^{d\times T}$ and $\mathbf{Y} \in \mathbb{R}^{d\times S}$ be free, standard Gaussian matrices, and define

$$\mathbf{H}_0 = \frac{1}{T}\mathbf{X}\mathbf{X}^\top, \qquad \mathbf{M} = \frac{1}{S}\mathbf{Y}\mathbf{Y}^\top, \qquad d/T \to q \in (0,1), \qquad d/S \to \psi \in (0,1).$$

Theorem (Spectrum for Sample Covariance Preconditioning, S. *et al.*, '26b, [8])

Let

$$\overline{\mathbf{H}}_0 = (1-\psi)\mathbf{M}^{-1/2}\mathbf{H}_0\mathbf{M}^{-1/2}, \qquad p = \frac{\psi}{1-\psi}.$$

Then the empirical spectral distribution of $\overline{\mathbf{H}}_0$ *converges almost surely to a compact law with support*

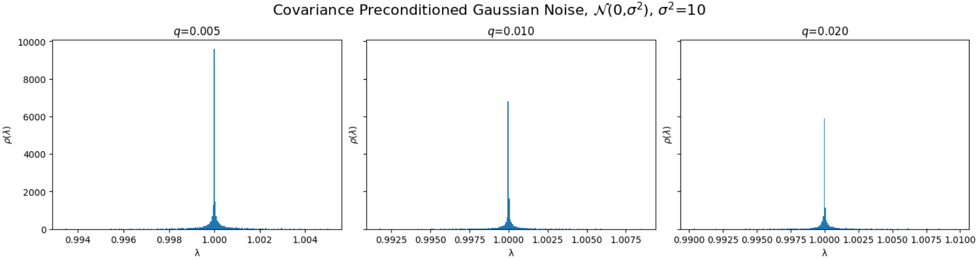
$$[\lambda_-, \lambda_+], \qquad \lambda_\pm = 1 + q + 2p \pm 2\sqrt{(1+p)(q+p)}. \tag{6}$$

How does preconditioning affect the noise?

Given this model $\mathbf{H} = \mathbf{H}_0 + \mathbf{H}_1$, the goal is to precondition the bulk $\mathbf{H}_0$ without amplifying the noise $\mathbf{H}_1$.

$$\mathbf{M}^{-1/2} \left(\mathbf{H}_0 + \mathbf{H}_1\right) \mathbf{M}^{-1/2} \approx \mathbf{\Pi} + \mathbf{M}_0^{-1/2}\mathbf{H}_1\mathbf{M}_0^{-1/2}$$

where $\mathbf{\Pi}$ is a projector matrix.



Full covariance inversion can widen the preconditioned spectrum

Let

$$\mathbf{M} = \frac{1}{S}\mathbf{Y}\mathbf{Y}^\top, \qquad d/S \to \psi \in (0,1), \qquad \mathbf{\Sigma} = \sigma_\xi^2 \mathbf{I},$$

Then

$$\frac{1}{d}\operatorname{tr}\left(\mathbf{M}^{-1}\mathbf{\Sigma}\right) \to \frac{\sigma_\xi^2}{1-\psi}, \qquad \lambda_{\max}(\mathbf{M}^{-1}) \to \left(1-\sqrt{\psi}\right)^{-2} \tag{7}$$

Remark

As $\psi \nearrow 1$, the upper edge grows. Thus uncontrolled full inversion can broaden the effective curvature bulk.

Table of Contents

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

1. enhancing local conditioning—decreasing $\hat{L}/\hat{c}$ ($\hat{L}/\hat{\mu}_{\mathrm{PL}}$ for nonconvexity) $\rightarrow$ speeds up contraction,

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

1. enhancing local conditioning—decreasing $\hat{L}/\hat{c}$ ($\hat{L}/\hat{\mu}_{\mathrm{PL}}$ for nonconvexity) $\rightarrow$ speeds up contraction,
2. reducing the preconditioned noise level K, lowering the effective noise floor,

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

1. enhancing local conditioning—decreasing $\hat{L}/\hat{c}$ ($\hat{L}/\hat{\mu}_{\mathrm{PL}}$ for nonconvexity) $\rightarrow$ speeds up contraction,
2. reducing the preconditioned noise level K, lowering the effective noise floor,
3. permitting a larger basin radius r, which jointly improves basin stability,

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

1. enhancing local conditioning—decreasing $\hat{L}/\hat{c}$ ($\hat{L}/\hat{\mu}_{\mathrm{PL}}$ for nonconvexity) $\rightarrow$ speeds up contraction,
2. reducing the preconditioned noise level K, lowering the effective noise floor,
3. permitting a larger basin radius r, which jointly improves basin stability,
4. responding robustly to Gaussian or heavy-tailed noise, and

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

1. enhancing local conditioning—decreasing $\hat{L}/\hat{c}$ ($\hat{L}/\hat{\mu}_{\mathrm{PL}}$ for nonconvexity) $\rightarrow$ speeds up contraction,
2. reducing the preconditioned noise level K, lowering the effective noise floor,
3. permitting a larger basin radius r, which jointly improves basin stability,
4. responding robustly to Gaussian or heavy-tailed noise, and
5. targeting the recoverable signal spikes.

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

1. enhancing local conditioning—decreasing $\hat{L}/\hat{c}$ ($\hat{L}/\hat{\mu}_{\mathrm{PL}}$ for nonconvexity) $\to$ speeds up contraction,
2. reducing the preconditioned noise level K, lowering the effective noise floor,
3. permitting a larger basin radius r, which jointly improves basin stability,
4. responding robustly to Gaussian or heavy-tailed noise, and
5. targeting the recoverable signal spikes.

Future Work: Generalization of Preconditioners

- This work was based on late stage convergence for Scientific Machine Learning, where generalization is straightforward.

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

1. enhancing local conditioning—decreasing $\hat{L}/\hat{c}$ ($\hat{L}/\hat{\mu}_{\mathrm{PL}}$ for nonconvexity) $\rightarrow$ speeds up contraction,
2. reducing the preconditioned noise level K, lowering the effective noise floor,
3. permitting a larger basin radius r, which jointly improves basin stability,
4. responding robustly to Gaussian or heavy-tailed noise, and
5. targeting the recoverable signal spikes.

Future Work: Generalization of Preconditioners

- This work was based on late stage convergence for Scientific Machine Learning, where generalization is straightforward.
- There are other machine learning frameworks such as LLMs, which have different goals for generalization.

Conclusion and Future Work

Conclusion:
A well-designed preconditioner $\mathbf{M}$ does the following:

1. enhancing local conditioning—decreasing $\hat{L}/\hat{c}$ ($\hat{L}/\hat{\mu}_{\mathrm{PL}}$ for nonconvexity) $\rightarrow$ speeds up contraction,
2. reducing the preconditioned noise level K, lowering the effective noise floor,
3. permitting a larger basin radius r, which jointly improves basin stability,
4. responding robustly to Gaussian or heavy-tailed noise, and
5. targeting the recoverable signal spikes.

Future Work: Generalization of Preconditioners

- This work was based on late stage convergence for Scientific Machine Learning, where generalization is straightforward.
- There are other machine learning frameworks such as LLMs, which have different goals for generalization.
- We aim to have a parameter *a priori* that informs the generalization in terms of a regret bound.

Questions?

Thank You!

Read the Paper!

References

[1] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization.* Jan. 29, 2017. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980[cs]. URL: http://arxiv.org/abs/1412.6980 (visited on 09/13/2024).

[2] Shunichi Amari. "Natural Gradient Works Efficiently in Learning". In: *Neural Computation* 10.2 (Feb. 1998), pp. 251–276. ISSN: 0899-7667. DOI: 10.1162/089976698300017746. URL: https://doi.org/10.1162/089976698300017746 (visited on 03/23/2026).

[3] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (1951). Publisher: Institute of Mathematical Statistics, pp. 400–407. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177729586. URL: https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-3/A-Stochastic-Approximation-Method/10.1214/aoms/1177729586.full (visited on 10/30/2024).

[4] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: *SIAM Review* 60.2 (2018), pp. 223–311. DOI: 10.1137/16M1080173. eprint: https://doi.org/10.1137/16M1080173. URL: https://doi.org/10.1137/16M1080173.

EMORY UNIVERSITY

References (cont.)

[5] Mitchell T. Scott et al. "Design Criteria for SGD Preconditioners: Local Conditioning, Noise Floors, and Basin Stability". In: *Transactions on Machine Learning Research* (2026). ISSN: 2835-8856. URL: https://openreview.net/forum?id=vo8FOBt6f6.

[6] Wenrui Hao et al. *Multiscale Neural Networks for Approximating Green's Functions*. 2024. arXiv: 2410.18439 [math.NA]. URL: https://arxiv.org/abs/2410.18439.

[7] Jeffrey Pennington and Yasaman Bahri. "Geometry of Neural Network Loss Surfaces via Random Matrix Theory". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 2798–2806. URL: https://proceedings.mlr.press/v70/pennington17a.html.

[8] Mitchell T. Scott et al. "Recoverable Spikes, Not Bulk Inversion: Random Matrix Theory of Low-Rank Preconditioning for Stochastic Optimization". In: (2026). Submitted to Conference.

Table of Contents

Spectrum of RMT Ensembles

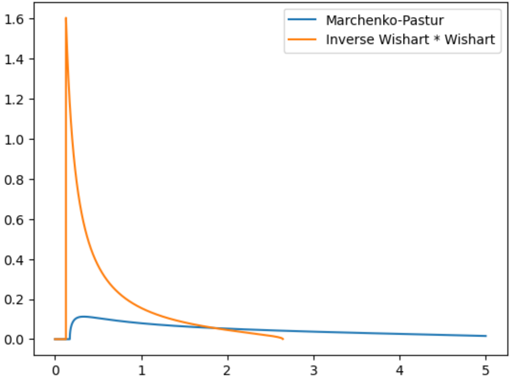


Figure 3: Comparing Marchenko - Pastur with $q = 2$ with Inverse Wishart Preconditioning with $q = 2, \psi = 0.5$