Reparameterization and Its Role in Optimization Dynamics

Hippolyte Labarrière, Cristian Vega, Cesare Molinari, Lorenzo Rosasco, Silvia Villa

4th UMI workshop "Mathematics for Artificial Intelligence and Machine Learning"

January 23, 2026

UniGe | DIBRIS | MaLGa MACHINE LEARNING GENOA CENTER

Context

Classical minimization task:

$$\min_{w \in \mathcal{W}} \mathcal{L}(w) \qquad (e.g.\ \mathcal{L}(w) = \frac{1}{2}\|Xw - y\|^2)$$

Context

Classical minimization task:

$$\min_{w \in \mathcal{W}} \mathcal{L}(w) \qquad (e.g.\ \mathcal{L}(w) = \frac{1}{2}\|Xw - y\|^2)$$

In most ML models (neural networks, LLMs, etc...) $\rightarrow$ Overparameterization

$$\min_{\theta \in \Theta} \mathcal{L}\left(h(\theta)\right), \quad \dim \Theta \gg \dim \mathcal{W}. \tag{1}$$

Context

Classical minimization task:

$$\min_{w \in \mathcal{W}} \mathcal{L}(w) \qquad (e.g.\ \mathcal{L}(w) = \frac{1}{2}\|Xw - y\|^2)$$

In most ML models (neural networks, LLMs, etc...) $\rightarrow$ Overparameterization

$$\min_{\theta \in \Theta} \mathcal{L}\left(h(\theta)\right), \quad \dim \Theta \gg \dim \mathcal{W}. \tag{1}$$

$\rightarrow$ Why is it efficient?

$\rightarrow$ Why overparameterization helps generalization?

Reparameterization

Idea: Study the effect of reparameterization on the optimization process

Original problem: | **Reparametrized problem:**

$\min_w \mathcal{L}(w)$ ⟶ $\min_\theta \mathcal{L}(h(\theta))$

↓

What happens in w? ⟵ Algorithm on θ

Gradient Flow vs Mirror flow

$$\min_{x \in \mathcal{X}} f(x)$$

Gradient Flow:

$$\frac{d}{dt}x(t) + \nabla f(x(t)) = 0, \; x(0) = x_0$$

$\rightarrow$ continuous version of Gradient Descent

Gradient Flow vs Mirror flow

$$\min_{x \in \mathcal{X}} f(x)$$

Gradient Flow:

$$\frac{d}{dt}x(t) + \nabla f(x(t)) = 0, \; x(0) = x_0$$

$\rightarrow$ continuous version of Gradient Descent

Mirror Flow (Alvarez et al., '04): for some convex and differentiable R,

$$\frac{d}{dt}\nabla R(x(t)) + \nabla f(x(t)) = 0, \; x(0) = x_0.$$

$\rightarrow$ modify the geometry of the space! (back to Gradient Flow for $R(x) = \frac{1}{2}\|x\|^2$)

Implicit Bias

By modifying the geometry of the space, **Mirror Flow** induces a different **implicit bias** from **Gradient Flow**.

Implicit Bias

By modifying the geometry of the space, **Mirror Flow** induces a different **implicit bias** from **Gradient Flow**.

Let $f(x) = \frac{1}{2}\|Ax - y\|^2$

- **Gradient Flow:** Converges towards

$$x_\infty = \arg\min\{\|x - x_0\|_2 : Ax = y\}$$

- **Mirror Flow:** Converges towards

$$\begin{aligned} x_\infty &= \arg\min\left\{D_R(x, x_0) : Ax = y\right\} \\ &= \arg\min\left\{R(x) - \langle \nabla R(x_0), x - x_0 \rangle : Ax = y\right\} \end{aligned}$$

Implicit Bias

By modifying the geometry of the space, **Mirror Flow** induces a different **implicit bias** from **Gradient Flow**.

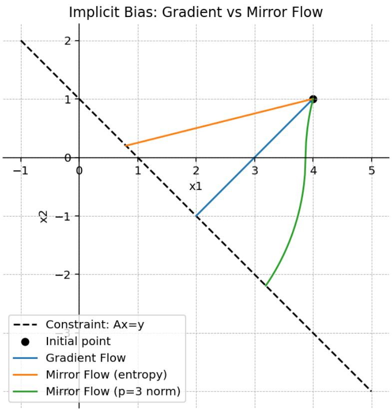
Let $f(x) = \frac{1}{2}\|Ax - y\|^2$

- **Gradient Flow:** Converges towards

$$x_\infty = \arg\min\{\|x - x_0\|_2 : Ax = y\}$$

- **Mirror Flow:** Converges towards

$$x_\infty = \arg\min\left\{D_R(x, x_0) : Ax = y\right\}$$



Implicit Bias

By modifying the geometry of the space, **Mirror Flow** induces a different **implicit bias** from **Gradient Flow**.

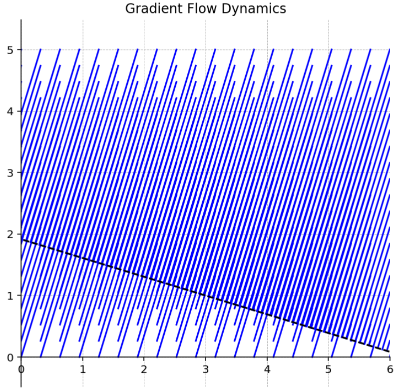
Let $f(x) = \frac{1}{2}\|Ax - y\|^2$

- **Gradient Flow:** Converges towards

$$x_\infty = \arg\min\{\|x - x_0\|_2 : Ax = y\}$$

- **Mirror Flow:** Converges towards

$$x_\infty = \arg\min\left\{D_R(x, x_0) : Ax = y\right\}$$



Implicit Bias

By modifying the geometry of the space, **Mirror Flow** induces a different **implicit bias** from **Gradient Flow**.

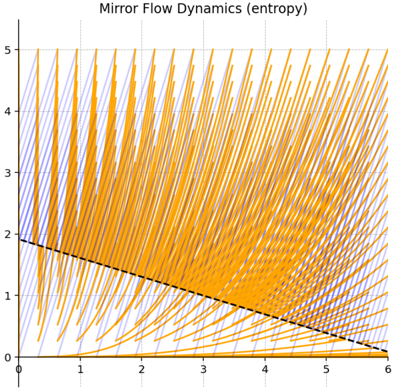
Let $f(x) = \frac{1}{2}\|Ax - y\|^2$

- **Gradient Flow:** Converges towards

$$x_\infty = \arg\min\{\|x - x_0\|_2 : Ax = y\}$$

- **Mirror Flow:** Converges towards

$$x_\infty = \arg\min\left\{D_R(x, x_0) : Ax = y\right\}$$



Implicit Bias

By modifying the geometry of the space, **Mirror Flow** induces a different **implicit bias** from **Gradient Flow**.

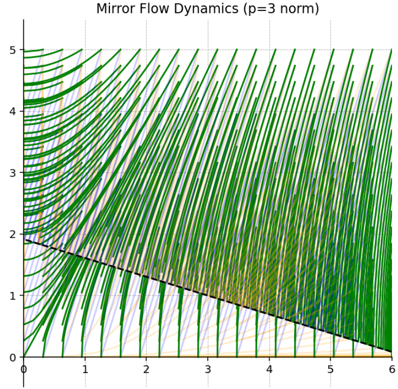
Let $f(x) = \frac{1}{2}\|Ax - y\|^2$

- **Gradient Flow:** Converges towards

$$x_\infty = \arg\min\{\|x - x_0\|_2 : Ax = y\}$$

- **Mirror Flow:** Converges towards

$$x_\infty = \arg\min\left\{D_R(x, x_0) : Ax = y\right\}$$



Back to Reparameterization

Let's train θ with **Gradient Flow**:

Original problem:

Reparametrized problem:

$$\min_w \mathcal{L}(w) \longrightarrow \min_\theta \mathcal{L}(h(\theta))$$

$$\downarrow$$

$$\frac{d}{dt}\theta(t) + \nabla_\theta \mathcal{L}(h(\theta(t))) = 0$$

Back to Reparameterization

Let's train θ with **Gradient Flow**:

Original problem: $\min_w \mathcal{L}(w)$ $\longrightarrow$ **Reparametrized problem:** $\min_\theta \mathcal{L}(h(\theta))$

$$\downarrow$$

$$\frac{d}{dt}\theta(t) + \nabla_\theta \mathcal{L}(h(\theta(t))) = 0$$

By chain rule: since $w(t) = h(\theta(t))$,

$$\frac{d}{dt}w(t) = \mathcal{J}_h(\theta(t))\frac{d}{dt}\theta(t) = -\mathcal{J}_h(\theta(t))\nabla_\theta \mathcal{L}(h(\theta(t))) = -\mathcal{J}_h(\theta(t))\mathcal{J}_h(\theta(t))^\top \nabla_w \mathcal{L}(w(t))$$

Back to Reparameterization

Let's train θ with **Gradient Flow**:

Original problem: **Reparametrized problem:**

$$\min_w \mathcal{L}(w) \longrightarrow \min_\theta \mathcal{L}(h(\theta))$$

$$\downarrow$$

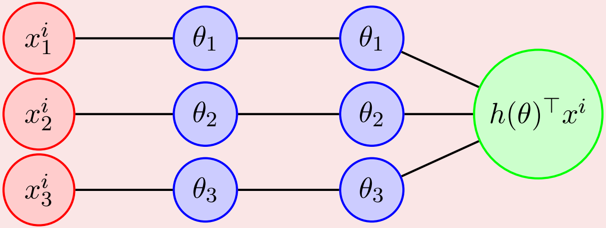
$$\frac{d}{dt} w(t) + \mathcal{J}_h(\theta(t))\mathcal{J}_h(\theta(t))^\top \nabla_w \mathcal{L}(w(t)) = 0 \longleftarrow \frac{d}{dt}\theta(t) + \nabla_\theta \mathcal{L}(h(\theta(t))) = 0$$

Is it a **Mirror Flow** in w?

$\rightarrow$ Yes, if $\mathcal{J}_h(\theta)\mathcal{J}_h(\theta)^\top = \nabla^2 R(w)^{-1}$ for some R!

Examples

Square reparameterization (Woodworth et al, '20):
Let $h(\theta) = \frac{1}{2}\theta \odot \theta$. Suppose $\mathcal{L}(w) = \frac{1}{2}\|Xw - y\|^2$.



Examples

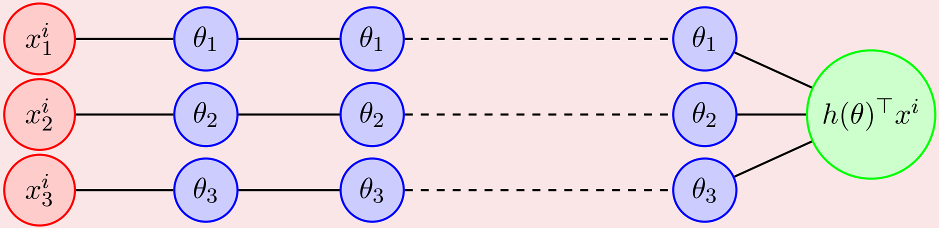
Square reparameterization (Woodworth et al, '20):
Let $h(\theta) = \frac{1}{2}\theta \odot \theta$.

$$\frac{d}{dt}w(t) + \underbrace{\theta(t) \odot \theta(t) \odot \nabla\mathcal{L}(w(t))}_{=\text{diag}(2w(t))\nabla\mathcal{L}(w(t))} = 0$$

$\rightarrow$ **Mirror Flow** with $R(w) = \frac{1}{2}\sum_{i=1}^{d} w_i \log(w_i) - w_i$.

Examples

Polynomial reparameterization (Woodworthet al, '20, Chou, Maly, Rauhut, '21):
Let $h(\theta) = \theta^{\odot L}$ for $L > 2$. Suppose $\mathcal{L}(w) = \frac{1}{2}\|Xw - y\|^2$.



Examples

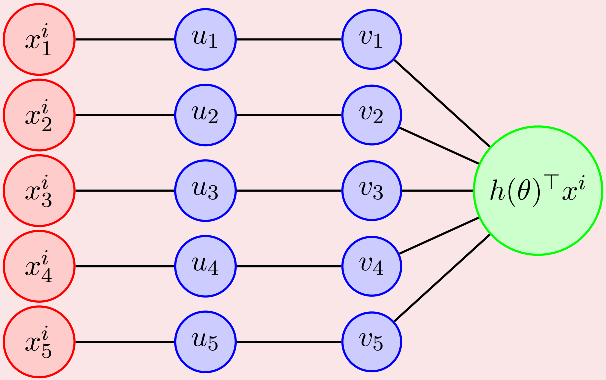
Polynomial reparameterization (Woodworthet al, '20, Chou, Maly, Rauhut, '21):
Let $h(\theta) = \theta^{\odot L}$ for $L > 2$.

$$\frac{d}{dt}w(t) + Lw(t)^{\odot(L-1)} \odot \nabla\mathcal{L}(w(t)) = 0.$$

$\rightarrow$ **Mirror Flow** with $R(w) = \langle\theta(0)^{L-2}, w\rangle - \frac{L}{2}\left\langle \mathbf{1}, w^{\frac{2}{L}} \right\rangle$.

Diagonal Linear Networks

Diagonal Linear Networks (Woodworth et al, '20, Moroshko et al., '20):
Let $h(\theta) = \frac{1}{2} u \odot v$ with $\theta = (u, v)$. Suppose $\mathcal{L}(w) = \frac{1}{2}\|Xw - y\|^2$.



Diagonal Linear Networks

Diagonal Linear Networks (Woodworth et al, '20, Moroshko et al., '20):
Let $h(\theta) = \frac{1}{2} u \odot v$ with $\theta = (u, v)$.

- (Woodworth et al., '20, Moroshko et al., '20) **Mirror Flow** in w with Mirror map:

$$R(w) = \frac{1}{2} \sum_{i=1}^{d} \left(2w_i \text{arcsinh}\left(\frac{2w_i}{\Delta_0}\right) - \sqrt{4w_i^2 + \Delta_0^2} + \Delta_0 \right) - \frac{1}{2} \left\langle \log \left| \frac{\theta_+(0)}{\theta_-(0)} \right|, \theta \right\rangle$$

- (Pesme et al., '21, Even et al., '23) **Stochasticity** helps generalization.
- (Nacson et al., '22) **Larger step-sizes** in Gradient Descent induce sparsity.
- (Papazov et al., '24) Adding **momentum** also helps generalization.

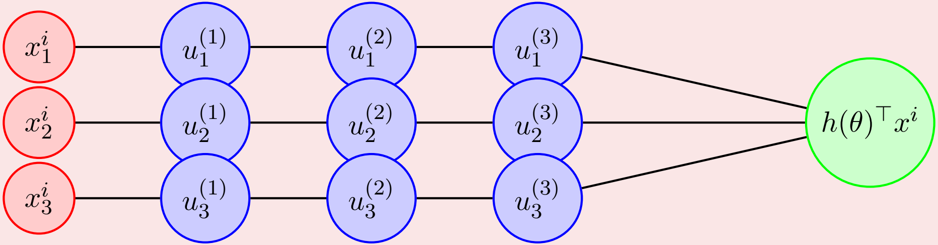
Diagonal Linear Networks

Diagonal Linear Networks (Woodworth et al, '20, Moroshko et al., '20):
Let $h(\theta) = \frac{1}{2} u \odot v$ with $\theta = (u, v)$.

- (Woodworth et al., '20, Moroshko et al., '20) **Mirror Flow** with an entropy map behaving as
  - $D_R(w, w_0) \sim \|w\|_1$ for small initialization $\to$ **encourages sparsity**.
  - $D_R(w, w_0) \sim \frac{1}{2}\|w - w_0\|_2^2$ for large initialization.
- (Pesme et al., '21, Even et al., '23) **Stochasticity** helps generalization.
- (Nacson et al., '22) **Larger step-sizes** in Gradient Descent induce sparsity.
- (Papazov et al., '24) Adding **momentum** also helps generalization.

Deep Diagonal Linear Networks

Deep Diagonal Linear Networks (Yun et al., '21, L. et al., '24):
Let $h(\theta) = \bigodot_{l=1}^{L} u^{(l)}$ with $\theta = (u^{(1)}, \ldots, u^{(L)})$. Suppose $\mathcal{L}(w) = \frac{1}{2}\|Xw - y\|^2$.



Deep Diagonal Linear Networks

Deep Diagonal Linear Networks (Yun et al., '21, L. et al., '24):
Let $h(\theta) = \bigodot_{l=1}^{L} u^{(l)}$ with $\theta = (u^{(1)}, \ldots, u^{(L)})$.

- (L. et al., '24) Under mild initialization assumptions,

  Gradient Flow in $\theta \equiv$ **Mirror Flow** in $w = h(\theta)$.

- (Yun et al., '21) For some structure of initialization,
  - Small initialization $\rightarrow \ell_1$ bias (sparsity).
  - More layers $\rightarrow$ stronger sparsity.

Other models and challenges

Matrix factorization (Gunasekar et al., '17, '18):
Let $h(\theta) = UV^\top$ with $\theta = (U, V)$.

For small initialization, $w(t)$ goes to the minimal nuclear norm solution.

Weight normalization (Salimans, Kingma, '16, Chou et al., '24):
Let $h(\theta) = g\frac{v}{\|v\|}$ with $\theta = (g, v)$.

$\rightarrow$ Sparsity inducing

Other models and challenges

Generalizing Mirror Flow (Vega et al., in preparation):

Some reparameterizations do not lead to Mirror Flow! But we can still characterize the implicit bias.

Studying Conservative Laws (Marcotte, Peyré, Gribonval, '23,'24,'25)

What are the invariant quantities during training?

Conclusion

Takeaways:

- Implicit bias of overparameterized models can be studied via optimization dynamics,
- Simple models give insights on more complex ones.

Limitations:

- Oversimplified models,
- Not adapted to non-linear models, i.e. $f_\theta(x^i) \neq \theta^\top x^i$,
- Challenging computations.

What about convergence to a global minimum?

$\rightarrow$ Oymak et al., '18, Chizat et al., '19, Li et al., '22, Chatterjee, '22, Kachaiev et al., in preparation.

Thank you for your attention!

Questions?

Related works:

- Vega. C., Molinari, C., Villa, S., Rosasco, L. (in preparation). Learning from data via over-parametrization.
- Labarrière, H., Molinari, C., Rosasco, L., Villa, S., Vega, C. (2024). Optimization Insights into Deep Diagonal Linear Networks. arXiv preprint arXiv:2412.16765.
- Kachaiev, O., Labarrière, H., Molinari, C., Villa, S. (in preparation). Geometric conditions for convergence of Gradient Flow to a global minimum.

My Website:

`https://hippolytelbrrr.github.io/pages/index_eng.html`

References I

- Alvarez, Bolte, Brahic. Hessian Riemannian gradient flows in convex programming, SICON, 2004.
- Woodworth et al, Kernel and rich regimes in overparametrized models, COLT, 2020.
- Chou, Maly, Rauhut, More is less: inducing sparsity via overparameterization, Information and Inference, 2021.
- Moroshko et al, Implicit bias in deep linear classification: Initialization scale vs training accuracy, NEURIPS, 2020.
- Pesme, Pillaud-Vivien, Flammarion, Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity, NEURIPS, 2021.
- Even, Pesme, Gunasekar, Flammarion, (S)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability, NEURIPS, 2023.
- Nacson, Ravichandran, Srebro, Soudry, Implicit bias of the step size in linear diagonal neural networks, ICML, 2022.
- Papazov, Pesme, Flammarion, Leveraging continuous time to understand momentum when training diagonal linear networks, AISTATS, 2024.
- Vega, Molinari, Villa, Rosasco, Learning from data via over-parametrization, in preparation.

References II

- Marcotte, Peyré, Gribonval, Abide by the law and follow the flow: Conservation laws for gradient flows, NEURIPS, 2023.
- Marcotte, Peyré, Gribonval, Keep the momentum: Conservation laws beyond euclidean gradient flows, arxiv, 2024.
- Marcotte, Peyré, Gribonval, Transformative or Conservative? Conservation laws for ResNets and Transformers, arxiv, 2025.
- Yun, Krishnan, Mobahi, A unifying view on implicit bias in training linear neural networks, ICLR, 2021.
- Gunasekar, Woodworth, Bhojanapalli, Neyshabur, Srebro, Characterizing implicit bias in terms of optimization geometry, NEURIPS, 2018.
- Gunasekar, Lee, Soudry, Srebro, Implicit regularization in matrix factorization: Implicit regularization in matrix factorization, ICML, 2018.
- Salimans, Kingma, Weight normalization: A simple reparameterization to accelerate training of deep neural networks, NEURIPS, 2016.
- Chou, Rauhut, Ward, Robust implicit regularization via weight normalization, Information and Inference, 2024.