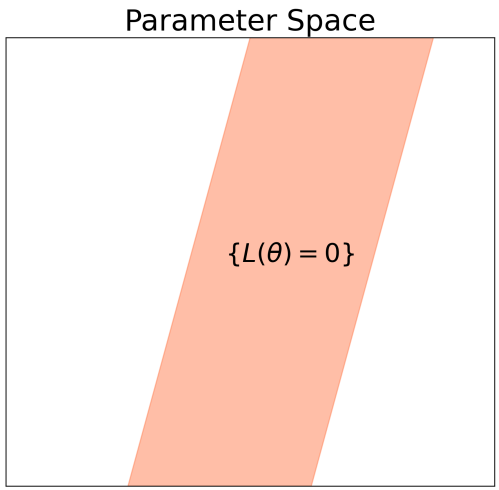
Implicit Bias of Mirror Flow for Shallow Neural Networks in Univariate Regression

Shuang Liang[†] Guido Montúfar[†,§]

[†] UCLA [§] MPI MIS

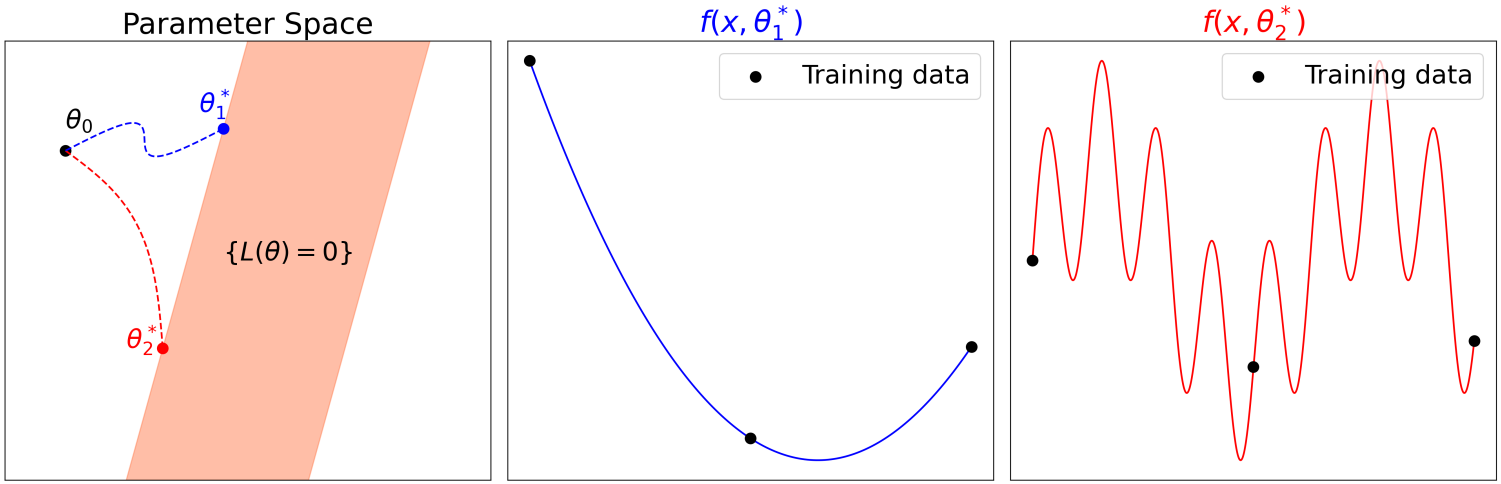
Implicit Bias of Optimization Procedures

- For overparametrized deep learning models, there may exist many candidate parameter values that minimize the empirical risk.



Implicit Bias of Optimization Procedures

- For overparametrized deep learning models, there may exist many candidate parameter values that minimize the empirical risk.



- Which specific solution is picked by the training algorithm?

Mirror Descent

- Let θ denote the model's parameter and L denote the loss function defined on the parameter space Θ.
- Gradient descent iteration:

$$\theta(t+1) = \arg\min_{\theta}\{\eta\langle\theta - \theta(t), \nabla L(\theta(t))\rangle + \frac{1}{2}\|\theta - \theta(t)\|^2\}$$

Mirror Descent

- Let θ denote the model's parameter and L denote the loss function defined on the parameter space Θ.
- Gradient descent iteration:

$$\theta(t+1) = \arg\min_{\theta}\{\eta\langle\theta - \theta(t), \nabla L(\theta(t))\rangle + \frac{1}{2}\|\theta - \theta(t)\|^2\}$$

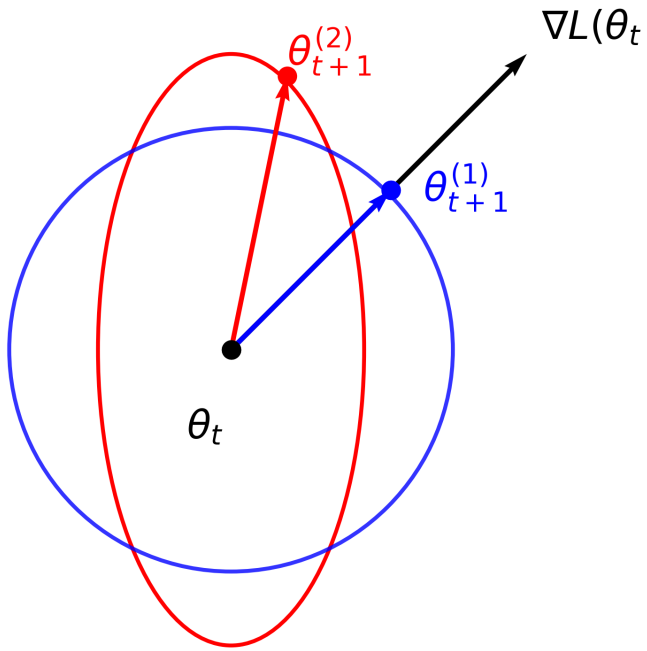
- Mirror descent iteration:

$$\theta(t+1) = \arg\min_{\theta}\{\eta\langle\theta - \theta(t), \nabla L(\theta(t))\rangle + D_\Phi(\theta, \theta(t))\}$$

Here $D_\Phi(\theta, \theta') = \Phi(\theta) - \Phi(\theta') - \langle\nabla\Phi(\theta'), \theta - \theta'\rangle$ is the ***Bregman divergence*** induced by the ***potential function*** $\Phi\colon \Theta \to \mathbb{R}$.

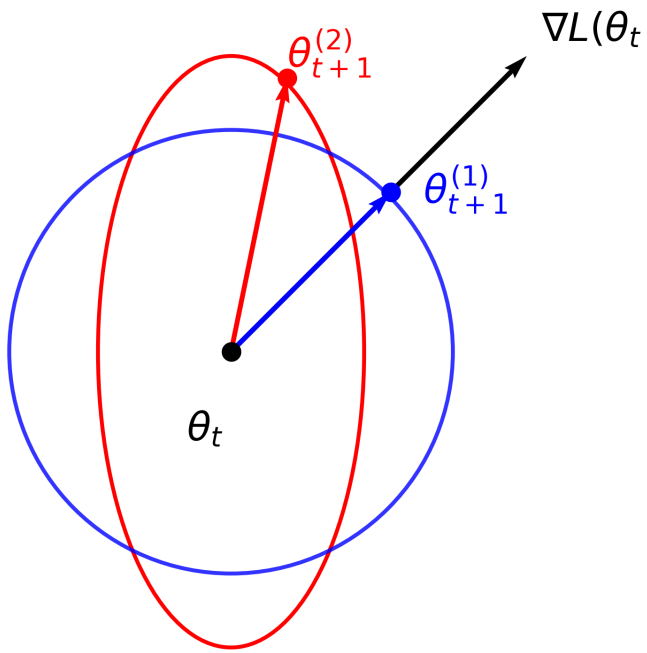
A Question

- Different notions of the ”distance” can lead to different update steps:



A Question

- Different notions of the ”distance” can lead to different update steps:



- How does the choice of the “distance” affect the learned network function?

Set-up

- Consider shallow neural networks with n hidden units, d input units and activation function σ:

$$f(x, \theta) = \sum_{j=1}^{n} \Big(a_j \sigma(\langle x, w_j \rangle - b_j) \Big) + d, \quad \forall x \in \mathbb{R}^d.$$

- Assume the parameter is randomly initialized by i.i.d. samples:

$$\forall j,\ w_j \sim \mathcal{W} = \mathsf{Unif}(\mathbb{S}^{d-1}),\ b_j \sim \mathcal{B},\ a_j \sim \frac{1}{\sqrt{n}} \mathcal{A}; \quad d \sim \mathcal{D},$$

  where $\mathcal{B}$, $\mathcal{A}$ and $\mathcal{D}$ are sub-Gaussian random variables. Assume $\mathcal{B}$ has continuous density function $p_{\mathcal{B}}(\cdot)$, and $(\mathcal{W}, \mathcal{B}) \overset{d}{=} (-\mathcal{W}, -\mathcal{B})$.

Set-up

- Consider shallow neural networks with n hidden units, d input units and activation function σ:

$$f(x,\theta)=\sum_{j=1}^{n}\Big(a_j\sigma(\langle x,w_j\rangle-b_j)\Big)+d,\quad \forall x\in\mathbb{R}^d.$$

- Assume the parameter is randomly initialized by i.i.d. samples:

$$\forall j,\ w_j\sim\mathcal{W}=\mathbf{Unif}(\mathbb{S}^{d-1}),\ b_j\sim\mathcal{B},\ a_j\sim\frac{1}{\sqrt{n}}\mathcal{A};\quad d\sim\mathcal{D},$$

  where $\mathcal{B}$, $\mathcal{A}$ and $\mathcal{D}$ are sub-Gaussian random variables. Assume $\mathcal{B}$ has continuous density function $p_{\mathcal{B}}(\cdot)$, and $(\mathcal{W},\mathcal{B})\overset{d}{=}(-\mathcal{W},-\mathcal{B})$.

- Consider training data set $\{(x_i,y_i)\}_{i=1}^m$ that satisfies $x_i\neq x_j$ when $i\neq j$ and $\{\|x_i\|_2\}_{i=1}^m\subset\mathbf{supp}(p_{\mathcal{B}})$, and the mean squared error

$$L(\theta)=\sum_{i=1}^{m}(f(x_i,\theta)-y_i)^2.$$

Set-up

- With infinitesimal learning rate η, the training dynamics of mirror descent is captured by ***mirror flow***:

$$\frac{\mathrm{d}}{\mathrm{dt}}\theta(t) = -\eta\Big(\nabla^2\Phi(\theta(t))\Big)^{-1}\nabla L(\theta(t)).$$

Set-up

- With infinitesimal learning rate η, the training dynamics of mirror descent is captured by ***mirror flow***:

$$\frac{\mathrm{d}}{\mathrm{dt}}\theta(t) = -\eta\Big(\nabla^2\Phi(\theta(t))\Big)^{-1}\nabla L(\theta(t)).$$

- Let $\boldsymbol{y} = [y_1, \ldots, y_m]^T \in \mathbb{R}^m$, $\boldsymbol{f}(t) = [f(x_1, \theta(t)), \ldots, f(x_m, \theta(t))]^T \in \mathbb{R}^m$, and $J_\theta \boldsymbol{f}(t)$ denote the Jacobian matrix of $\boldsymbol{f}(t)$ with respect to θ. By the chain rule,

$$\frac{\mathrm{d}}{\mathrm{dt}}\boldsymbol{f}(t) = -n\eta H(t)(\boldsymbol{f}(t) - \boldsymbol{y}),$$
$$H(t) \triangleq n^{-1} J_\theta \boldsymbol{f}(t)(\nabla^2\Phi(\theta(t)))^{-1} J_\theta(\boldsymbol{f}(t))^T.$$

- The evolution of $\boldsymbol{f}(t)$ is governed by the kernel-like matrix $H(t)$.

Unscaled Potential Function

Assumption (Unscaled potential)

Assume the potential function $\Phi\colon \Theta \to \mathbb{R}$ satisfies:

- Φ is separable, i.e., $\Phi(\theta) = \sum_{j=1}^{\dim(\Theta)} \phi(\theta_j - \hat{\theta}_j)$ for some real-valued function ϕ, where $\hat{\theta}_j$ is selected as the initialized value of θ_j;
- Φ is twice continuously differentiable;
- $\nabla^2\Phi(\theta)$ is positive definite for all $\theta \in \Theta$.

- Examples of unscaled potentials:
  - $\phi(x) = x^2$, which recovers gradient flow;
  - $\phi(x) = x^p + \omega x^2$, for $\omega > 0$ and $p > 2$;
  - $\phi(x) = \cosh(x)$.

Main Result I

Theorem (Implicit bias of mirror flow with unscaled potential; Part I)

For shallow NN with ReLU activation, $d \geq 1$ input units, and sufficiently large width n, consider mirror flow with unscaled potential Φ and learning rate $\eta = \Theta(1/n)$. There exist constants $C_1, C_2 > 0$ such that with high probability over random parameter initialization, for any $t \geq 0$,

$$\|\theta(t) - \theta(0)\|_\infty \leq C_1 n^{-\frac{1}{2}}, \ \lim_{n\to\infty} \|H(t) - H(0)\|_2 = 0, \ \|\boldsymbol{f}(t) - \boldsymbol{y}\|_2 \leq e^{-\eta_0 C_2 t} \|\boldsymbol{f}(0) - \boldsymbol{y}\|_2.$$

Main Result I

Theorem (Implicit bias of mirror flow with unscaled potential; Part I)

For shallow NN with ReLU activation, $d \geq 1$ input units, and sufficiently large width n, consider mirror flow with unscaled potential Φ and learning rate $\eta = \Theta(1/n)$. There exist constants $C_1, C_2 > 0$ such that with high probability over random parameter initialization, for any $t \geq 0$,

$$\|\theta(t) - \theta(0)\|_\infty \leq C_1 n^{-\frac{1}{2}}, \ \lim_{n\to\infty} \|H(t) - H(0)\|_2 = 0, \ \|\boldsymbol{f}(t) - \boldsymbol{y}\|_2 \leq e^{-\eta_0 C_2 t} \|\boldsymbol{f}(0) - \boldsymbol{y}\|_2.$$

Corollary

Let $\theta(\infty) = \lim_{t\to\infty} \theta(t)$ and $\theta_{\mathrm{GF}}(\infty)$ denote the limiting point of gradient flow on the same training data and initial parameter. For any given $x \in \mathbb{R}^d$,

$$\lim_{n\to\infty} |f(x, \theta(\infty)) - f(x, \theta_{\mathrm{GF}}(\infty))| = 0.$$

- For many model with *fixed* parameter dimension, the implicit biases of mirror flow and gradient flow are different [1, 2].

Theorem (Implicit bias of mirror flow with unscaled potential; Part II)

Assume $d = 1$. Let μ denote the measure associated with $(\mathcal{W}, \mathcal{B})$. For any given $x \in \mathbb{R}$, we have that $\lim_{n\to\infty} |f(x, \theta(\infty)) - f(x, \theta(0)) - \bar{h}(x)| = 0$, where $\bar{h}(\cdot)$ is the solution to the following variational problem:

$$\min_{h \in \mathcal{F}_{\mathrm{ReLU}}} \mathcal{G}_1(h) + \mathcal{G}_2(h) + \mathcal{G}_3(h) \quad \text{s.t. } h(x_i) = y_i - f(x_i, \theta(0)),\ \forall i \in [m],$$

$$\begin{cases} \mathcal{G}_1(h) = \displaystyle\int_{\mathrm{supp}(p_{\mathcal{B}})} \frac{(h''(x))^2}{p_{\mathcal{B}}(x)} \mathrm{d}x \\ \mathcal{G}_2(h) = \Big(\lim_{x\to+\infty} h'(x) + \lim_{x\to-\infty} h'(x) \Big)^2 \\ \mathcal{G}_3(h) = \dfrac{1}{\mathbb{E}[\mathcal{B}^2]} \Big(\displaystyle\int_{\mathrm{supp}(p_{\mathcal{B}})} h''(x)|x| \mathrm{d}x - 2h(0) \Big)^2. \end{cases}$$

Here
$\mathcal{F}_{\mathrm{ReLU}} = \Big\{ h(x) = \int \alpha(w, b)[wx - b]_+ \mathrm{d}\mu \colon \alpha \text{ uniformly continuous on } \mathrm{supp}(\mu) \Big\}$.

Theorem (Implicit bias of mirror flow with unscaled potential; Part II)

Assume $d = 1$. Let μ denote the measure associated with $(\mathcal{W}, \mathcal{B})$. For any given $x \in \mathbb{R}$, we have that $\lim_{n\to\infty} |f(x, \theta(\infty)) - f(x, \theta(0)) - \bar{h}(x)| = 0$, where $\bar{h}(\cdot)$ is the solution to the following variational problem:

$$\min_{h \in \mathcal{F}_{\mathrm{ReLU}}} \mathcal{G}_1(h) + \mathcal{G}_2(h) + \mathcal{G}_3(h) \quad \text{s.t. } h(x_i) = y_i - f(x_i, \theta(0)),\ \forall i \in [m],$$

$$\begin{cases} \mathcal{G}_1(h) = \displaystyle\int_{\mathrm{supp}(p_{\mathcal{B}})} \frac{(h''(x))^2}{p_{\mathcal{B}}(x)} \mathrm{d}x & \text{(favors smoothness)} \\ \mathcal{G}_2(h) = \Big(\lim_{x\to+\infty} h'(x) + \lim_{x\to-\infty} h'(x) \Big)^2 \\ \mathcal{G}_3(h) = \dfrac{1}{\mathbb{E}[\mathcal{B}^2]} \Big(\displaystyle\int_{\mathrm{supp}(p_{\mathcal{B}})} h''(x)|x| \mathrm{d}x - 2h(0) \Big)^2. \end{cases}$$

Here
$\mathcal{F}_{\mathrm{ReLU}} = \Big\{ h(x) = \int \alpha(w, b)[wx - b]_+ \mathrm{d}\mu \colon \alpha \text{ uniformly continuous on } \mathrm{supp}(\mu) \Big\}$.

Theorem (Implicit bias of mirror flow with unscaled potential; Part II)

Assume $d = 1$. Let μ denote the measure associated with $(\mathcal{W}, \mathcal{B})$. For any given $x \in \mathbb{R}$, we have that $\lim_{n\to\infty} |f(x, \theta(\infty)) - f(x, \theta(0)) - \bar{h}(x)| = 0$, where $\bar{h}(\cdot)$ is the solution to the following variational problem:

$$\min_{h \in \mathcal{F}_{\mathrm{ReLU}}} \mathcal{G}_1(h) + \mathcal{G}_2(h) + \mathcal{G}_3(h) \quad \text{s.t. } h(x_i) = y_i - f(x_i, \theta(0)),\ \forall i \in [m],$$

$$\begin{cases} \mathcal{G}_1(h) = \displaystyle\int_{\mathrm{supp}(p_{\mathcal{B}})} \frac{(h''(x))^2}{p_{\mathcal{B}}(x)} \mathrm{d}x & \text{(favors smoothness)} \\ \mathcal{G}_2(h) = \Big(\lim_{x\to+\infty} h'(x) + \lim_{x\to-\infty} h'(x) \Big)^2 & \text{(favors inverse slopes at } \pm\infty) \\ \mathcal{G}_3(h) = \dfrac{1}{\mathbb{E}[\mathcal{B}^2]} \Big(\displaystyle\int_{\mathrm{supp}(p_{\mathcal{B}})} h''(x)|x| \mathrm{d}x - 2h(0) \Big)^2. \end{cases}$$

Here
$\mathcal{F}_{\mathrm{ReLU}} = \Big\{ h(x) = \int \alpha(w, b)[wx - b]_+ \mathrm{d}\mu \colon \alpha \text{ uniformly continuous on } \mathrm{supp}(\mu) \Big\}$.

Main Result I

- If $\mathcal{B}$ is supported on $[-B, B]$ for some $B > 0$, then

$$\mathcal{G}_3(h) = \frac{1}{\mathbb{E}[\mathcal{B}^2]} \Big(B\big(\lim_{x \to +\infty} h'(x) - \lim_{x \to -\infty} h'(x) \big) - \big(h(B) + h(-B) \big) \Big)^2.$$

  - If h is above the x-axis, then $\mathcal{G}_3$ promotes $h'(+\infty) > h'(-\infty)$ and thus favors a U-shaped function.
  - If h is below, $\mathcal{G}_3$ favors an inverted U-shape.

Experiments

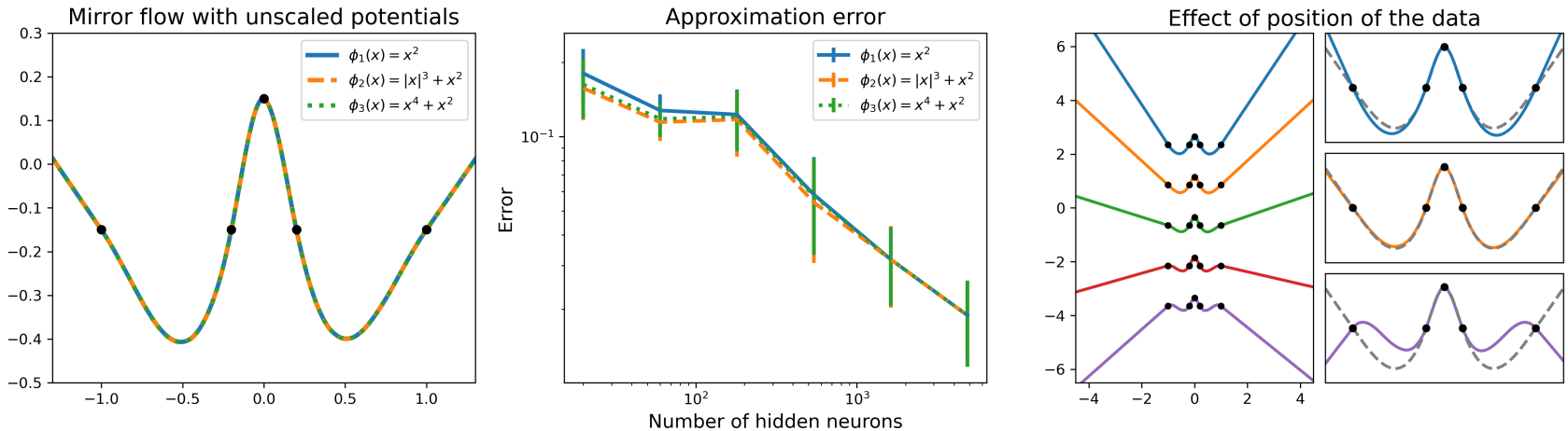


Figure 1: Left: ReLU networks trained with mirror flow on a common data set using unscaled potentials: $\phi_1 = x^2$, $\phi_2 = |x|^3 + x^2$, and $\phi_3 = x^4 + x^2$. Middle: L^∞-error between the solution to the variational problem and the trained networks, against the network width. Right: ReLU networks trained with gradient descent on five different data sets with shifted labels.

Scaled Potential Function

Can we design training algorithms capable of inducing implicit biases depending on the parameter geometry?

Scaled Potential Function

Can we design training algorithms capable of inducing implicit biases depending on the parameter geometry?

Assumption (Scaled potential)

Recall that n denotes the network width. Assume the potential function $\Phi\colon \Theta \to \mathbb{R}$ satisfies:

- Φ can be written in the form of $\Phi(\theta) = \frac{1}{n^2}\sum_{k=1}^{p}\phi\big(n(\theta_k - \hat{\theta}_k)\big)$ for some real-valued function ϕ, where $\hat{\theta}_j$ is selected as the initialized value of θ_j;
- ϕ takes the form of $\phi(x) = \frac{1}{1+\omega}\left(\psi(x) + \omega x^2\right)$, where $\omega > 0$ and $\psi \in C^3(\mathbb{R})$ is a convex function on $\mathbb{R}$.

Main Result II

Theorem (Implicit bias of mirror flow with scaled potential)

Assume $\mathcal{B}$ is compactly supported. For shallow NN with absolute value activation, $d \geq 1$ input units and sufficiently large width n, consider mirror flow with a scaled potential Φ with sufficiently large parameter w, and learning rate $\eta = \Theta(1/n)$. There exist constants $C_1, C_2 > 0$ such that with high probability over the random parameter initialization, for any $t \geq 0$,

$$\|\theta(t) - \theta(0)\|_\infty \leq C_1 n^{-1}, \ \|f(t) - y\|_2 \leq e^{-\eta_0 C_2 t} \|f(0) - y\|_2.$$

Theorem (Implicit bias of mirror flow with scaled potential)

Assume $\mathcal{B}$ is compactly supported. For shallow NN with absolute value activation, $d \geq 1$ input units and sufficiently large width n, consider mirror flow with a scaled potential Φ with sufficiently large parameter w, and learning rate $\eta = \Theta(1/n)$. There exist constants $C_1, C_2 > 0$ such that with high probability over the random parameter initialization, for any $t \geq 0$,

$$\|\theta(t) - \theta(0)\|_\infty \leq C_1 n^{-1}, \ \|\boldsymbol{f}(t) - \boldsymbol{y}\|_2 \leq e^{-\eta_0 C_2 t} \|\boldsymbol{f}(0) - \boldsymbol{y}\|_2.$$

Assume univariate input data $d = 1$. For any given $x \in \mathbb{R}$ that $\lim_{n\to\infty} |f(x, \theta(\infty)) - f(x, \theta(0)) - \bar{h}(x)| = 0$, where $\bar{h}(\cdot)$ solves:

$$\min_{h \in \mathcal{F}_{\mathrm{Abs}}} \int_{\mathrm{supp}(p_{\mathcal{B}})} D_\phi\Big(\frac{h''(x)}{2p_{\mathcal{B}}(x)}, 0\Big) p_{\mathcal{B}}(x) \mathrm{d}x \quad \text{s.t. } h(x_i) = y_i - f(x_i, \theta(0)), \ \forall i \in [m].$$

Here D_ϕ denotes the Bregman divergence on $\mathbb{R}$ induced by ϕ and $\mathcal{F}_{\mathrm{Abs}} = \{h(x) = \int \alpha(w, b)|wx - b| \mathrm{d}\mu \colon \alpha \text{ is even and uniformly continuous on } \mathrm{supp}(\mu)\}$.

Main Result II

$$\min_{h \in \mathcal{F}_{\mathrm{Abs}}} \int_{\mathsf{supp}(p_{\mathcal{B}})} D_{\phi}\left(\frac{h''(x)}{2p_{\mathcal{B}}(x)}, 0\right) p_{\mathcal{B}}(x)\mathrm{d}x \quad \text{s.t. } h(x_i) = y_i - f(x_i, \theta(0)),\ \forall i \in [m].$$

- The parameter initialization $p_{\mathcal{B}}$ determines the strength of the penalization of the function's second derivative at different locations;
- The scaled potential ϕ determines the strength of the penalization of the different magnitudes of the second derivative.

Experiments

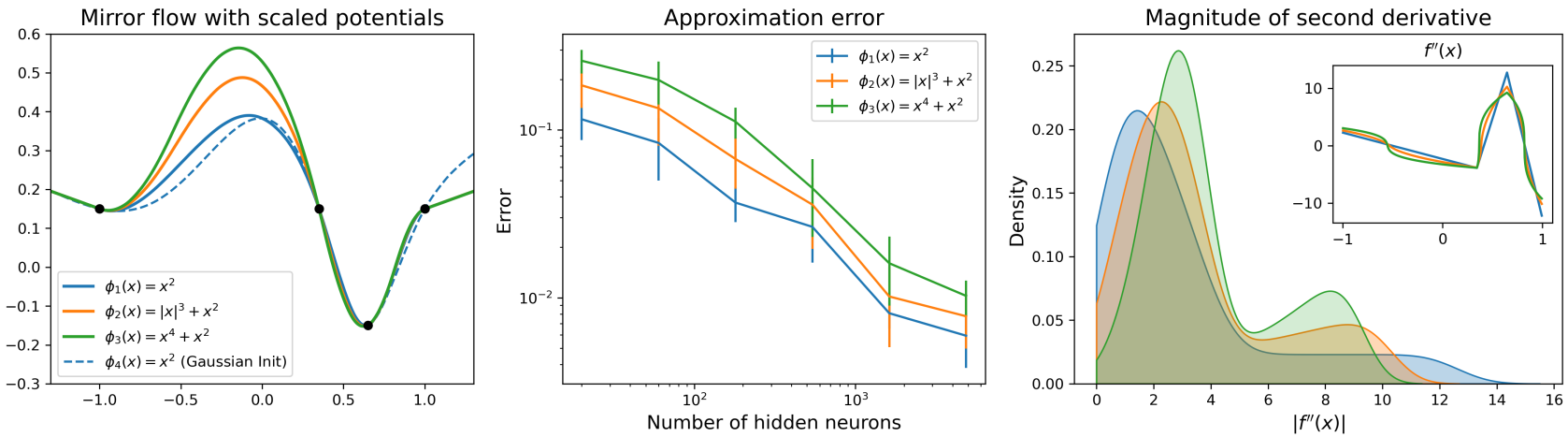


Figure 2: Left: absolute value networks trained with mirror flow on a common data set using scaled potentials: $\phi_1 = x^2$, $\phi_2 = |x|^3 + x^2$, and $\phi_3 = x^4 + x^2$. Middle: L^∞-error between the solution to the variational problem and the trained networks, against the network width. Right: distribution of the magnitude of the second derivative of the solutions to the variational problems.

References i

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro.
Characterizing implicit bias in terms of optimization geometry.
In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1832–1841. PMLR, 2018.

H. Sun, K. Gatmiry, K. Ahn, and N. Azizan.
A unified approach to controlling implicit regularization via mirror descent.
Journal of Machine Learning Research, 24(393):1–58, 2023.