Math of ML : Exercises 10 *

November 24, 2025

Exercise 1 (Reparametrized gradient descent as mirror descent)**.** *Let $X \in \mathbb{R}^{n\times d}$ be the design matrix, $y \in \mathbb{R}^n$ be the response variable, and $R(w) = \frac{1}{n}\|Xw - y\|_2^2$ be the empirical risk, where $w \in \mathbb{R}^d$. We consider the setting where $d > n$ and X is of full rank; in particular, there exists an infinite number of minimizers of R. In what follows, $1_d \in \mathbb{R}^d$ denotes a vector with all entries equal to one and for any function $f : \mathbb{R} \to \mathbb{R}$ and any vector $v \in \mathbb{R}^d$ we denote by $f(v)$ the vector whose j-th entry is equal to $f(v_j)$ (e.g., v^2 denotes the vector whose j-th entry is v_j^2). Finally, the notation $u > 0$ means that each component of u is strictly greater than 0.*

1. *Consider the gradient flow dynamics $\frac{d}{dt}w_t = -\nabla R(w_t)$ with the initialization $w_0 = 0$. Let $w_\infty = \lim_{t\to\infty} w_t$. What can you say about w_∞ (does it converge to a minimizer of R? If so, what special property does w_∞ satisfy among the infinite set of possible minimizers of R?)*

2. *We now turn to a reparametrized gradient flow. One drawback of the parametrization $w_t = u_t^2$ analyzed in Lecture 10 is that it does not allow us to represent vectors w_t with negative coordinates. To fix this, in the rest of this question we consider the parametrization*

$$w_t = u_t^2 - v_t^2,$$

where $v_t, u_t \in \mathbb{R}^d$. Let $z_t = (u_t, v_t)^\top \in \mathbb{R}^{2d}$ and for $z = (u, v)^\top \in \mathbb{R}^{2d}$ define $\tilde{R}(z) = \frac{1}{4}R(u^2 - v^2)$. Consider the gradient flow dynamics $\frac{d}{dt}z_t = -\nabla\tilde{R}(z_t)$. Compute $\frac{d}{dt}w_t$. ***Hint:*** *your answer should take the form $-A_t\nabla R(w_t)$ for some diagonal matrix A_t that depends only on u_t^2 and v_t^2.*

3. *Prove that for all $t \geq 0$ the dynamics of the previous question satisfy*

$$\frac{d}{dt}\left[\left(u_t^2 + v_t^2\right)^2 - \left(u_t^2 - v_t^2\right)^2\right] = 0.$$

4. *Assume that the initialization (u_0, v_0) satisfies $u_0^2 > 0$ and $v_0^2 > 0$. Combining your answers to the two previous parts, show that w_t follows the mirror descent flow*

$$\frac{d}{dt}w_t = -\left(\nabla^2\psi_{u_0,v_0}(w_t)\right)^{-1}\nabla R(w_t)$$

for some mirror map ψ_{u_0,v_0} that you should identify (here $\nabla^2\psi_{u_0,v_0}(w_t)$ denotes the Hessian matrix of ψ_{u_0,v_0} at the point w_t).

***Hint: first verify that for any** $c > 0$ **we have** $\frac{d^2}{dx^2}(x\,\mathrm{arcsinh}(x/c) - \sqrt{x^2 + c^2}) = 1/\sqrt{c^2 + x^2}$.*

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

5. *Consider the initialization*

$$u_0 = v_0 = \alpha 1_d,$$

where $\alpha \in \mathbb{R}$. *Let* $(w_t(\alpha))_{t\geq 0}$ *denote the mirror flow dynamics identified in the previous question and let* $w_\infty(\alpha) = \lim_{t\to\infty} w_t(\alpha)$. *What can you say about*

(a) $\lim_{\alpha\to 0} w_\infty(\alpha)$*;*

(b) $\lim_{\alpha\to\infty} w_\infty(\alpha)$*?*

Solution 1.

1. *The optimization objective* $w \mapsto R(w)$ *is convex. Hence, the gradient flow dynamics converges to a minimizer. By the assumptions of this question, there exists an infinite number of minimizers. However, since* $\frac{d}{dt}w_t = -\frac{2}{n}X^\top(Xw_t - y)$, *we have for all* $t \geq 0$, $\frac{d}{dt}w_t \in \mathrm{range}(X^\top)$. *Because* $w_0 = 0$, *it follows that* $w_\infty \in \mathrm{range}(X^\top)$, *which by the results of Lecture* 9 *implies that* w_∞ *is a minimum* ℓ_2 *norm minimizer of the objective* R.

2. *For vectors* $a, b \in \mathbb{R}^d$, *let* $a \odot b$ *denote their Hadamard product (i.e.,* $(a \odot b)_j = a_j b_j$*). We have*

$$\begin{aligned}
\frac{d}{dt}u_t &= -\frac{1}{4}\frac{2}{n}X^\top(Xw_t - y) \odot (2u_t) = -\frac{1}{n}X^\top(Xw_t - y) \odot u_t, \\
\frac{d}{dt}v_t &= -\frac{1}{4}\frac{2}{n}X^\top(Xw_t - y) \odot (-2v_t) = \frac{1}{n}X^\top(Xw_t - y) \odot v_t.
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{d}{dt}w_t &= \frac{d}{dt}(u_t^2 - v_t^2) \\
&= \frac{d}{dt}(u_t^2 - v_t^2) \\
&= 2u_t \odot \frac{d}{dt}u_t - 2v_t \odot \frac{d}{dt}v_t \\
&= -\frac{2}{n}X^\top(Xw_t - y) \odot u_t^2 - \frac{2}{n}X^\top(Xw_t - y) \odot v_t^2 \\
&= -\nabla R(w_t) \odot u_t^2 - \nabla R(w_t) \odot v_t^2 \\
&= -\mathrm{diag}(u_t^2 + v_t^2)\nabla R(w_t), \qquad (1)
\end{aligned}$$

where $\mathrm{diag}(v)$ *denotes a diagonal matrix with a vector* v *on its diagonal.*

3. *By the computations in the previous question, we have*

$$\begin{aligned}
\frac{d}{dt}\left[\left(u_t^2 + v_t^2\right)^2 - \left(u_t^2 - v_t^2\right)^2\right] &= \frac{d}{dt}\left[\left(u_t^2 + v_t^2\right)^2 - \left(u_t^2 - v_t^2\right)^2\right] \\
&= 4\frac{d}{dt}u_t^2 v_t^2 \\
&= 8u_t\left[\frac{d}{dt}u_t\right]v_t^2 + 8v_t\left[\frac{d}{dt}v_t\right]u_t^2 \\
&= 0.
\end{aligned}$$

4. *By the previous part, for all* $t \geq 0$ *we have*

$$(u_t^2 + v_t^2)^2 = (u_t^2 - v_t^2)^2 + 4u_0^2 v_0^2 = w_t^2 + 4u_0^2 v_0^2.$$

Hence, by (1) we have

$$\frac{d}{dt}w_t = -\text{diag}\left(\sqrt{w_t^2 + 4u_0^2 v_0^2}\right)\nabla R(w_t),$$

To show that the above is a mirror decent flow, we need to find a strictly-convex function ψ such that

$$\left(\nabla^2\psi(w_t)\right)^{-1} = \text{diag}\left(\sqrt{w_t^2 + 4u_0^2 v_0^2}\right).$$

The diagonal structure suggests to look for ψ of the form:

$$\psi(w) = \sum_{i=1}^{d}\phi_i(w_i).$$

Hence, we need to solve the equation:

$$\phi_i''(w_i) = \frac{1}{\sqrt{w_i^2 + 4(u_0)_i^2(v_0)_i^2}}.$$

By the hint, the above has a solution

$$\phi_i(w_i) = w_i\text{arcsinh}\left(\frac{w_i}{2(u_0)_i(v_0)_i}\right) - \sqrt{w_i^2 + 4(u_0)_i^2(v_0)_i^2},$$

which completes the proof.

5. *With the specified initialization, by the previous question we know that $(w_t)_{t\geq 0}$ follows mirror decent dynamics with the mirror map*

$$\psi_\alpha(w) = \sum_{i=1}^{d} w_i\text{arcsinh}\left(\frac{w_i}{2\alpha^2}\right) - \sqrt{w_i^2 + 4\alpha^4},$$

with the associated Bregman divergence

$$D_{\psi_\alpha}(w, w') = \sum_{i=1}^{d} w_i\left(\text{arcsinh}\left(\frac{w_i}{2\alpha^2}\right) - \text{arcsinh}\left(\frac{w_i'}{2\alpha^2}\right)\right) - \sum_{i=1}^{d}\sqrt{w_i^2 + 4\alpha^4} + \sum_{i=1}^{d}\sqrt{(w')_i^2 + 4\alpha^4}.$$

In particular, letting $w' = w_0 = 0$ we have

$$D_{\psi_\alpha}(w, w_0) = \sum_{i=1}^{d} w_i\text{arcsinh}\left(\frac{w_i}{2\alpha^2}\right) - \sum_{i=1}^{d}\sqrt{w_i^2 + 4\alpha^4} + 2d\alpha^2. \tag{2}$$

We know from lecture 10 that mirror descent flow converges to a minimizer that minimizes the Bregman divergence $D_{\psi_\alpha}(w, w_0)$. Using this fact we can understand the limiting cases of the implicit bias as follows.

   (a) *As $\alpha \to 0$, the Bregman divergence (2) becomes proportional to $\|w\|_1$ (the first term in (2) becomes dominant). Thus, as $\alpha \to 0$, the mirror descent flow $(w_t)_{t\to 0}$ converges to a minimum ℓ_1 norm solution.*

(b) As $\alpha \to \infty$ we have

$$\begin{aligned}
&x \operatorname{arcsinh}\left(\frac{x}{2\alpha^2}\right) - \sqrt{x^2+4\alpha^4} \\
&= x \log\left(\sqrt{\frac{x^2}{4\alpha^4}+1} + \frac{x}{2\alpha^2}\right) - \sqrt{x^2+4\alpha^4} \\
&\approx x \cdot \frac{x}{2\alpha^2} - \sqrt{x^2+4\alpha^4} \\
&= \frac{1}{2\alpha^2}\left(x^2 - \sqrt{\frac{x^2}{4\alpha^4}+1}\right) \\
&\approx \frac{1}{2\alpha^2}\left(x^2-1\right).
\end{aligned}$$

Thus, as $\alpha \to \infty$, the mirror descent flow $(w_t)_{t\to 0}$ converges to a minimum ℓ_2 norm solution.

The above arguments can be made more rigorous. For a quantification of values of α so that either the ℓ_1 or ℓ_2 norm implicit bias behaviour approximately kicks in, see Woodworth, Gunasekar, Lee, Moroshko, Savarese, Golan, Soudry, and Srebro [2020, Theorem 2].

Exercise 2 ((Practical exercise) Verification of Results of Exercise 1)**.** *In this exercise, we numerically verify some of the findings of Exercise 1. For this exercise, you may initialize $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$ arbitrarily, as long as $d > n$ and X is of full rank. In what follows, we use the notation of Exercise 1.*

1. *Verify numerically that for small enough step size $\eta > 0$, the predictor $w_t = u_t^2 - v_t^2$ that evolves according to the gradient descent dynamics*

$$z_0 = (u_0, v_0) = \alpha 1_{2d}, \quad z_{t+1} = z_t - \eta \nabla \tilde{R}(z_t)$$

approximately matches the mirror descent dynamics $(w_t(\alpha))_{t\geq 0}$ defined by

$$w_0(\alpha) = 0, \quad \nabla\psi_\alpha(w_{t+1}(\alpha)) = \nabla\psi_\alpha(w_t(\alpha)) - \eta\nabla R(w_t(\alpha)), \tag{3}$$

where ψ_α denotes the mirror map $\psi_{u_0,v_0} = \psi_{\alpha 1_d, \alpha 1_d}$ that you identified in Exercise 1.

2. *For a list of initialization scale values α ranging from 10^{-10} to 10^3, simulate the mirror descent dynamics $w_t(\alpha)$ defined in (3) for a large enough number of iterations T_α so that $w_{T_\alpha}(\alpha) \approx w_\infty(\alpha)$ (i.e., run mirror descent until approximate convergence). For each α, plot the values of $\|w_{T_\alpha}(\alpha)\|_2 - \|w_2^*\|_2$ and $\|w_{T_\alpha}(\alpha)\|_1 - \|w_1^*\|_1$, where w_p^* denotes any minimum ℓ_p norm minimizer of R. Does the observed behaviour confirm your theoretical predictions made in Exercise 1.5?*

Solution 2. *For part 2 see Woodworth et al. [2020, Figure 1].*

Exercise 3 ((Practical exercise) Regularization vs optimization paths)**.** *Let $X \in \mathbb{R}^{100\times 2}$ be a matrix with i.i.d. $\mathcal{N}(0,1)$ entries (thus, $n = 100$ and $d = 2$). Let $Y = Xw^* + Z$, where $w^* = (1,0)^\top$ and $Z \sim \mathcal{N}(0, I_{100})$, where I_{100} denotes the 100×100 identity matrix. Let $R(w) = \frac{1}{n}\|Xw - Y\|_2^2$ and consider*

$$w_{1,\lambda} \in \operatorname{argmin}_{w\in\mathbb{R}^2} R(w) + \lambda\|w\|_1 \quad \text{and} \quad w_{2,\lambda} \in \operatorname{argmin}_{w\in\mathbb{R}^2} R(w) + \lambda\|w\|_2^2.$$

2. *Choose a set of parameter values* Λ_2 *such that* $0 \in \Lambda_2$ *and the maximum value of* $\lambda \in \Lambda_2$ *is large enough so that* $w_{2,\lambda} \approx 0$. *Plot the regularization path* $(w_{2,\lambda})_{\lambda \in \Lambda_2}$. *On the same graph, plot the optimization path* $(w_t^{\mathrm{gd}})_{t=0}^T$ *of gradient descent iterates* $w_0^{\mathrm{gd}} = 0; w_{t+1}^{\mathrm{gd}} = w_t^{\mathrm{gd}} - \eta \nabla R(w_t^{\mathrm{gd}})$, *where* $\eta > 0$ *is a small enough step size and* T *is a chosen number of iterations to ensure approximate convergence.*

3. *Choose a set of parameter values* Λ_1 *such that* $0 \in \lambda_1$ *and the maximum value of* $\lambda \in \Lambda_1$ *is large enough so that* $w_{1,\lambda} \approx 0$. *Plot the regularization path* $(w_{1,\lambda})_{\lambda \in \Lambda_1}$. *On the same graph, plot the optimization path* $(w_t^{\mathrm{md}})_{t=0}^T$ *of mirror descent iterates* $w_0^{\mathrm{md}} = 10^{-10}(1,1)^\top; \log(w_{t+1}^{\mathrm{md}}) = \log(w_t^{\mathrm{md}}) - \eta \nabla R(w_t^{\mathrm{md}})$, *where* $\eta > 0$ *is a small enough step size and* T *is a chosen number of iterations to ensure approximate convergence. Note: this mirror descent path is generated via the negative entropy mirror map* $\psi(w) = \sum_{i=1}^d w_i \log w_i - w_i$.

For both simulation setups above, think about the following points.

- *Does the optimization path nearly match the regularization path?*
- *If the paths do not match, which path is better? You may judge the quality of an optimization/regularization path by how close the best point on a given path gets to the optimal parameter* w^*.
- *Which path is computationally cheaper to generate?*

Remark 1. *For computing the regularization path* $(w_{1,\lambda})_{\lambda \in \Lambda_1}$ *search for a Lasso solver implementation in your favourite language. For example, in Python you may use* https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html.

References

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.