Mirror Descent and Implicit Bias

Gabriel Peyré
CNRS & ENS
École Normale Supérieure
gabriel.peyre@ens.fr

March 7, 2021

Abstract

These notes should serve as a tutorial on mirror-type descents and it also introduces the main ideas on their implicit bias.

1 Mirror Descent and Implicit Bias

1.1 Bregman Divergences

We consider a smooth strictly convex "entropy" function ψ such that $\|\nabla\psi(x)\|$ goes to $+\infty$ as $x \to \partial\,\mathrm{dom}(\psi)$. We denote

$$\psi^*(u) \overset{\text{def.}}{=} \sup_{x \in \mathrm{dom}(\psi)} \langle u,\, x\rangle - \psi(x)$$

its Legendre transform. In this case of "Legendre-type" entropy function, $\nabla\psi : \mathrm{dom}(\psi) \to \mathrm{dom}(\psi^\star)$ and $\nabla\psi^*$ are bijection reciprocal one from the other.

One then defines the associated Bregman divergence

$$\mathrm{D}_\psi(x|y) \triangleq \psi(x) - \psi(y) - \langle \nabla\psi(y),\, x - y\rangle.$$

It is positive, convex in x (but not necessarily in y), not necessarily symmetric, and "distance-like".

For $\psi = \|\cdot\|^2$ one has $\nabla\psi = \nabla\psi^* = \mathrm{Id}$, and one recovers the Euclidean distance. For $\psi_{\mathrm{KL}}(x) = \sum_i x_i \log(x_i) - x_i + 1$ one has $\nabla\psi = \log$ and $\nabla\psi^* = \exp$, and one obtains the relative entropy, also known as Kullback-Leibler

$$\mathrm{D}_{\psi_{\mathrm{KL}}}(x|y) = \sum_i x_i \log(x_i/y_i) - x_i + y_i.$$

When $\psi_{\mathrm{Burg}}(x) = \sum_i -\log(x_i) + x_i - 1$ on $\mathbb{R}_+^d$, $\nabla\psi_{\mathrm{Burg}}(x) = \nabla\psi^*(x) = -1/x$ and associated divergence

$$\mathrm{D}_{\psi_{\mathrm{Burg}}}(x|y) = \sum_i -\log(y_i/x_i) - x_i/y_i + 1. \tag{1}$$

These examples can be generalized to power entropies

$$\psi_\alpha(x) \triangleq \sum_i \frac{|x_i|^\alpha - \alpha(x_i - 1) - 1}{\alpha(\alpha - 1)} \tag{2}$$

with special cases

$$\psi_1(x) \triangleq \psi_{\mathrm{KL}} = \sum_i x_i \log(x_i) - x_i + 1 \quad \text{and} \quad \psi_0(x) \triangleq \psi_{\mathrm{Burg}} = \sum_i -\log(x_i) + x_i - 1.$$

They are defined on $\mathbb{R}^d$ if $\alpha > 1$ and $\mathbb{R}_+^d$ if $\alpha \leqslant 1$.

Remark 1 (Matricial divergences)*.* Given an entropy function $\psi_0(x)$ on vectors $x \in \mathbb{R}^d$ which is invariant under permutation of the indices, one lifts it to symmetric matrices $X \in \mathbb{R}^{d\times d}$ as

$$\psi(X) \triangleq \psi_0(\Lambda(X)) \quad \text{where} \quad X = U_X \operatorname{diag}(\Lambda(X)) U_X^\top$$

is the eigen-decomposition of X, where $\Lambda(X) = (\lambda_i(X))_{i=1}^d \in \mathbb{R}^d$ are the eigenvalues. Typically, if $\psi_0(x) = \sum_i h(x_i)$ then $\psi(X) = \operatorname{tr}(h(X))$ where h is extended to matrices as $h(X) \triangleq U_X \operatorname{diag}(h(\lambda_i(X)))U_X^\top$. If ψ_0 is convex and smooth, so is ψ, and

$$\nabla\psi(X) = U_X \operatorname{diag}(\nabla\psi_0(\Lambda(X)))U_X^\top.$$

For instance, if $h(s) = s\log(x) - s + 1$ is the Shannon entropy, this defines the quantum Shannon entropy as

$$\mathrm{D}_\psi(X) = \operatorname{tr}(X\log(X) - X\log(Y) - X + Y)$$

and if $h(s) = -\log(s)$ then $\mathrm{D}_\psi(X) = -\log\det(X)$.

Remark 2 (Cizard divergences)*.* When defined on $\mathbb{R}_+^d$, these divergence should not be confounded with Cizar divergences which reads

$$\mathrm{C}_\psi(x|y) \stackrel{\text{def.}}{=} \sum_i y_i\psi(x_i/y_i) + \psi'_\infty \sum_{y_i=0} x_i,$$

which are jointly convex in x and y. Only for $\psi = \psi_{\mathrm{KL}}$ one has $\mathrm{D}_{\psi_{\mathrm{KL}}} = \mathrm{C}_{\psi_{\mathrm{KL}}}$.

1.2 Mirror descent

We consider the following implicit stepping

$$x_{k+1} = \underset{x\in\operatorname{dom}(\psi)}{\operatorname{argmin}}\ f(x) + \frac{1}{\tau}\mathrm{D}_\psi(x|x_k).$$

Its explicit version then reads by Taylor expanding f at x_k

$$\begin{aligned} x_{k+1} &= \underset{x\in\operatorname{dom}(\psi)}{\operatorname{argmin}}\ f(x_k) + \langle x - x_k,\, \nabla f(x_k)\rangle + \frac{1}{\tau}\mathrm{D}_\psi(x|x_k), \\ &= \underset{x\in\operatorname{dom}(\psi)}{\operatorname{argmin}}\ \langle x,\, \nabla f(x_k)\rangle + \frac{1}{\tau}\mathrm{D}_\psi(x|x_k). \end{aligned}$$

The fact that ψ is Legendre type allows to ignore the constraint, and the solution satisfies the following first order condition

$$\nabla f(x_k) + 1/\tau[\nabla\psi(x_{k+1}) - \nabla\psi(x_k)] = 0$$

so that it can be explicitly computed

$$x_{k+1} = (\nabla\psi^*)[\nabla\psi(x_k) - \tau\nabla f(x_k)] \tag{3}$$

For $\psi = \|\cdot\|^2/2$ one recovers the usual Euclidean gradient descent. For $\psi(x) = \sum_i x_i\log(x_i)$, this defines the multiplicative updates

$$x_{k+1} = x_k \odot \exp(-\tau\nabla f(x_k))$$

where $\odot$ is the entry-wise multiplication of vectors.

Note that introducing the "dual" variable $u_k \triangleq \nabla\psi(x_k)$, one has

$$u_{k+1} = u_k - \tau h(u_k) \quad \text{where} \quad h(u) \triangleq \nabla f(\nabla\psi^*(u)). \tag{4}$$

Note however that in general h is not a gradient field, so this is not in general a gradient flow.

Mirror flow. When $\tau \to 0$, one obtains the following expansion

$$x_{k+1} = (\nabla\psi^*)[\nabla\psi(x_k)] - \tau[\partial^2\psi^*](\nabla\psi(x_k)) \times \nabla f(x_k) + o(\tau)$$

so that defining $x(t) = x_k$ for $t = k\tau$ the limit is the following flow

$$\dot{x}(t) = -H(x(t))\nabla f(x(t)) \quad \text{where} \quad H(x) \triangleq [\partial^2\psi^*](\nabla\psi(x)) = [\partial^2\psi(x)]^{-1} \tag{5}$$

so that this is a gradient flow on a very particular type of manifold, of "Hessian type". Note that if $\psi = f$, then one recovers the flow associated to Newton's method.

Convergence. Convergence theory (ensuring convergence and rates) for mirror descent is the same as for the usual gradient descent, and one needs to consider relative L-smoothness, and if possible also relative μ-strong convexity,

$$\mu\,\mathrm{D}_\psi \leqslant \mathrm{D}_f \leqslant L\,\mathrm{D}_\psi \quad \Longleftrightarrow \quad \forall\, x,\ \mu\partial^2\psi(x) \leqslant \partial^2 f(x) \leqslant L\partial^2\psi(x).$$

If $L < +\infty$, then one has $f(x_k) - f(x^\star) \leqslant O(\mathrm{D}_\psi(x^\star|x_0)/k)$ while if both $0 < \mu \leqslant L < +\infty$, then $\mathrm{D}_\psi(x_k|x^\star) \leqslant O(\mathrm{D}_\psi(x^\star|x_0)(1-\mu/L)^k)$. The advantages of using Bregman geometry are two-fold: this can improves the conditioning μ/L (some function might be non-smooth for the Euclidean geometry but smooth for some Bregman geometry, and can avoid introducing constraint in the optimization problem) and this can also lower the radius of the domain $\mathrm{D}_\psi(x^\star|x_0)$. For instance, assuming the solution belongs to the simplex, and using $x_0 = \mathbb{1}_d/d$, then $\mathrm{D}_{\psi_{\mathrm{KL}}}(x^\star|x_0) \leqslant \log(d)$ whereas for the ℓ^2 Euclidean distance, one only has the bound $\|x^\star - x_0\|^2 \leqslant d$.

1.3 Re-parameterized flows

One can consider a change of variable $x = \varphi(z)$ where $\varphi : \mathbb{R}^p \mapsto \mathcal{X} \subset \mathbb{R}^d$ is a smooth map, and perform the gradient descent on the function $g(z) \triangleq f(\varphi(z))$. Then one has

$$\nabla g(z) = [\partial\varphi(z)]^\top \nabla f(x)$$

so that, denoting $z(t)$ the gradient flow $\dot{z} = -\nabla g(z)$ of g, and $x(t) \triangleq \varphi(z(t))$, one has $\dot{x}(t) = [\partial\varphi(z(t))]\dot{z}(t)$ and thus $x(t)$ solves the following equation

$$\dot{x} = -Q(z)\nabla f(x) \quad \text{with} \quad Q(z) \triangleq [\partial\varphi(z)][\partial\varphi(z)]^\top \in \mathcal{S}_+^{d\times d}$$

So unless φ is a bijection, this is not a gradient flow over the x variable. If φ is a bijection, then this is a gradient flow associated to the field of tensors ("manifold") $Q(\varphi^{-1}(x))$. The issue is that even in this case, in general H might fail to be a Hessian manifold, so this does not correspond to a mirror descent flow.

Dual parameterization If ψ is an entropy function, then the parametrization $x = \nabla\psi^*(z)$, i.e. $\varphi = \nabla\psi^*$, then $Q(z) = [\partial^2\psi^*(z)]^2$, i.e. $Q(\varphi^{-1}(x)) = [\partial^2\psi(x)]^{-2}$ is not of Hessian-type in general, but rather a squared-Hessian manifold. For instance, when $\psi^*(z) = \exp(z)$, then $Q(\varphi^{-1}(x)) = \mathrm{diag}(1/x_i^2)$, which surprisingly is the hessian metric associated to Burg's entropy $-\sum_i \log(x_i)$.

Example: power-type parameterization We consider power entropies (2), on $\mathbb{R}_+^d$, for $\alpha \leqslant 1$, for which

$$H(x) = [\partial^2\psi(x)]^{-1} \propto \mathrm{diag}(x_i^{2-\alpha}).$$

Remark than when using the parameterization $x = \varphi(z) = (z_i^b)_i$ then

$$Q(\varphi^{-1}(x)) = [\partial\varphi(z)][\partial\varphi(z)]^\top \propto \mathrm{diag}(z_i^{2(b-1)}) = \mathrm{diag}(x_i^{2(b-1)/b})$$

so if one selects $2(1-1/b) = 2-\alpha$ i.e. $2/b = \alpha$, the re-parameterized flow is equal to the flow on a Hessian manifold. For instance, when setting $b = 2$, $\alpha = 1$, i.e. using the parmeterization $x = z^2$, one retrieves the flow on the manifold for the Shannon entropy ("Fisher-Rao" geometry). Note that when $b \to +\infty$, one obtains $\alpha = 0$, i.e. the flow is the one of the Burg's entropy $\psi(x) = -\sum_i \log(x_i)$ (which we saw above as also being associated to the parameterization $x = \exp(z)$).

Counter-example: SDP matrices We now consider semi-definite symmetric matrices $X \in \mathcal{S}_+^{d\times d}$, together with the parameterization $X = \varphi(Z) = ZZ^\top$ for $Z \in \mathbb{R}^{d\times d}$. In this case, denoting $g(Z) = f(ZZ^\top)$, one has

$$\nabla g(Z) = [\nabla f(X) + \nabla f(X)^\top] Z$$

so that the flow $\dot Z = -\nabla g(Z)$ is equivalent to the following flow on symmetric (and it maintains positivity as well)

$$\dot X = X[\nabla_S f(X)] + [\nabla_S f(X)] X \tag{6}$$

where the symmetric gradient is

$$\nabla_S f(X) \triangleq [\nabla f(X)] + [\nabla f(X)]^\top$$

So most likely (6) cannot be written as a usual gradient flow on a manifold which would be a hessian of a convex function. To mimic the diagonal case (or vectorial case above), the most natural quantitate would have been the spectral entropy $\psi(X) \triangleq \operatorname{tr}(X\log(X) - X + \mathrm{Id})$, whose gradient is $\log(X)$, but there is no closed form expression for the derivative of the log unfortunately. Another simpler approach to mimic ψ_{-1} is to use $\psi(X) = -\operatorname{tr}(\log(X)) = -\log\det(X)$, because the Hessian and its inverse can be computed

$$\partial^2\psi(X) : S \mapsto -X^{-1} S X^{-1}.$$

1.4 Implicit Bias

We consider the problem

$$\min_{x\in\mathbb{R}^d} f(x) = L(Ax) \triangleq \sum_i \ell(\langle a_i,\, x\rangle, y_i),$$

where the loss is coercive such that $\ell(\cdot, y_i)$ has a unique minimizer at y_i. The typical example is $f(x) = \|Ax - y\|^2$ for $\ell(u,v) = (u-v)^2$. We do not impose that L is convex, and simply assumes convergence of the considered optimization method to the set of global minimizers. The set of global minimizers is thus the affine space

$$\operatorname{argmin} f = \{x \;;\; Ax = y\}.$$

The simplest optimization method is just gradient descent

$$x_{k+1} = x_k - \tau\nabla f(x_k) \quad \text{where} \quad \nabla f(x) = A^\top \nabla L(Ax).$$

As $\tau \to 0$, one defines $x(t) = x_k$ for $t = k\tau$ and consider the flow

$$\dot x(t) = -\nabla f(x(t)).$$

The implicit bias of the descent (and the flow) is given by the orthogonal projection.

Proposition 1. *If $x_k \to x^\star \in \operatorname{argmin} f$, then*

$$x^\star = \underset{x\in\operatorname{argmin} f}{\operatorname{argmin}} \|x - x_0\|.$$

The following Proposition, whose proof can be found in [1] generalizes this proposition to the case of an arbitrary mirror flow.

Proposition 2. *If x_k defined by (3) (resp. $x(t)$ defined by (5)) is such that x_k (resp. $x(t)$) converges to $x^\star \in \operatorname{argmin} f$, then*

$$x^\star = \underset{x\in\operatorname{argmin} f}{\operatorname{argmin}} \mathrm{D}_\psi(x|x_0). \tag{7}$$

Proof. From the dual variable evolution (4), since $\nabla f(x) \in \operatorname{Im}(A^\top)$, one has that $y_k - y_0 \in \operatorname{Im}(A^\top)$, so that in the limit

$$y^\star - y_0 = \nabla\psi(x^\star) - \nabla\psi(x_0) \in \operatorname{Im}(A^\top). \tag{8}$$

Note that $\nabla \mathrm{D}_\psi(x|x_0) = \nabla\psi(x) - \nabla\psi(x_0)$, and $\operatorname{Im}(A^\top) = \operatorname{Ker}(A)^\perp$ is the space orthogonal to $\operatorname{argmin} f$ so that (8) are the optimality conditions of the strictly convex problem (7). $\square$

In particular, for the Shannon entropy (equivalently when using the $x = z^2$ parameterization), as $x_0 \to 0$, by doing the expansion of $\mathrm{KL}(x|x_0)$ one has

$$x^\star \to \underset{x \in \operatorname{argmin} f, x \geqslant 0}{\operatorname{argmin}} \sum_i |\log((x_0)_i)| x_i,$$

which is a weighted ℓ^1 norm (so in particular it induces sparsity in the solution, it is a Lasso-type problem).

When using more general parameterizations of the form $x = z^b$ for $b > 0$, this corresponds to using the power entropy ψ_α for $\alpha = 2/b$, and one can check that the associated limit bias for small x_0 is still an ℓ^1, but with a different weighting scheme. For $x = \exp(z)$ (or $b \to +\infty$) one obtains Burg's entropy defined in (1) so that the limit bias is $\sum_i x_i/(x_0)_i$. The use of $x = z^2$ parameterization (which can be generalized to $x = u \odot v$ for signed vectors) was introduced in [2], and its associated implicit regularization is detailed in [3, 5]. It is possible to analyze this sparsity-inducing behavior in a quantitative way, see for instance [4, Thm.2] One can generalize this parameterization to arbitrary (not only positive vector) by using $x = u^2 - v^2$ or $x = u \odot v$ and the same type of bias appears, with now rather a (weighted) ℓ^1 norm.

References

[1] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[2] Peter D Hoff. Lasso, fractional norm and structured sparse estimation using a hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.

[3] Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *arXiv preprint arXiv:1909.05122*, 2019.

[4] 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.

[5] Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit regularization via hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*, 2019.