A Bregman firmly nonexpansive proximal operator for baryconvex optimization

Mastane Achab*
Deep Gambit Limited, Masdar City, Abu Dhabi, UAE
www.deepgambit.com

November 1, 2024

Abstract

We present a generalization of the proximal operator defined through a convex combination of convex objectives, where the coefficients are updated in a minimax fashion. We prove that this new operator is Bregman firmly nonexpansive with respect to a Bregman divergence that combines Euclidean and information geometries.

Notations The Euclidean norm of any vector $x \in \mathbb{R}^m$ ($m \geq 1$) is denoted $\|x\|$. For any integer $S \geq 1$, we denote by $\mathbf{1}_S$ the all-ones vector of size S and by Δ_S the probability simplex:

$$\Delta_S = \left\{ q = (q_1, \ldots, q_S) \in [0,1]^S : q_1 + \cdots + q_S = 1 \right\} \ .$$

The Kullback-Leibler divergence Kullback and Leibler [1951] will be denoted by "D_{KL}" throughout the paper: for any $q, r \in \Delta_S$, $D_{\mathrm{KL}}(r\|q) = \sum_{s=1}^S r_s \log\left(\frac{r_s}{q_s}\right)$. Let $h(q) = \sum_{s=1}^S q_s \log(q_s)$ be the negative entropy function defined over Δ_S; its gradient $\nabla h(q) = (1 + \log(q_s))_s$ with inverse $(\nabla h)^{-1}(u) = (\frac{e^{u_s - 1}}{\sum_{s'} e^{u_{s'} - 1}})_s$ for $u = (u_1, \ldots, u_S) \in \nabla h(\Delta_S)$. Given a differentiable function $\ell = (\ell_1, \ldots, \ell_S) : \mathbb{R}^m \to \mathbb{R}^S$, we denote by J_ℓ its Jacobian matrix. Finally, given $(x, q) \in \mathbb{R}^m \times \Delta_S$, we refer to the vector $J_\ell(x)^\intercal q = \sum_{s=1}^S q_s \nabla \ell_s(x)$ as the "q-barygradient of ℓ at x".

1 Problem statement

In this article we present a generalization of the convex optimization formalism (Boyd and Vandenberghe [2004]) that we call *baryconvex optimization* since it involves weighted convex objectives where the weights are learned in a minimax fashion.

*mastane@deepgambit.com

Definition 1 (Generalized proximal operator)**.** Let $\ell = (\ell_1, \dots, \ell_S) : \mathbb{R}^m \to \mathbb{R}^S$ for $m, S \geq 1$ and where ℓ_s is a differentiable convex function for each $s \in \{1, \dots, S\}$. Given $\lambda > 0$, we define our generalized proximal operator $\mathrm{prox}_{\lambda\ell}$ as follows: for all $(x, q) \in \mathbb{R}^m \times \Delta_S$

$$\mathrm{prox}_{\lambda\ell}(x, q) = \underset{(z,r)\in\mathbb{R}^m\times\Delta_S}{\arg\,\mathrm{minimax}} H_{x,q}(z, r) := r^\intercal \ell(z) + \frac{1}{2\lambda}\|z - x\|^2 - \frac{1}{\lambda} D_{\mathrm{KL}}(r\|q) \ .$$

First notice that for $S = 1$, the probability simplex is reduced to the singleton $\Delta_1 = \{1\}$, and we recover the standard proximal operator with a single convex loss function whose minimizers are exactly the fixed points of the prox. This paper proposes to extend well-known convex optimization methods such as the proximal point algorithm (PPA, see Rockafellar [1976]) and gradient descent (GD, see Boyd and Vandenberghe [2004]) to our general setting with $S \geq 1$.

> *Question: Can we compute a fixed point (if it exists) of the generalized prox in Definition 1?*

As will be shown, the answer provided by this paper is positive.

> *Answer: Yes, by leveraging a Bregman geometry that combines Euclidean and simplex structures.*

Saddle point We point out that the function $(z, r) \mapsto H_{x,q}(z, r)$ is strongly convex-concave (i.e. strongly convex in z and strongly concave in r, see e.g. Boyd and Vandenberghe [2004]) and, if $S \geq 2$, admits a unique saddle point $(x', q') = \mathrm{prox}_{\lambda\ell}(x, q)$ characterized by the stationarity condition $\nabla H_{x,q}(x', q') = 0$. Further, by the *minimax theorem*[1], we have:

$$\min_z \max_r H_{x,q}(z, r) = H_{x,q}(x', q') = \max_r \min_z H_{x,q}(z, r). \tag{1}$$

In the next sections, we propose to generalize some key components of the convex analysis toolbox (firm nonexpansion property Bauschke and Combettes [2011], PPA and GD methods) in order to find a fixed point of $\mathrm{prox}_{\lambda\ell}$ in the general case $S \geq 1$.

2 Bregman firm nonexpansiveness

We recall from Brohé and Tossings [2000]-Bauschke et al. [2003] that an operator T is Bregman firmly nonexpansive (BFNE) with respect to f if $\langle Tx - Ty, \nabla f(Tx) - \nabla f(Ty)\rangle \leq \langle Tx - Ty, \nabla f(x) - \nabla f(y)\rangle$, $\forall x, y$. Furthermore, if the BFNE operator has a fixed point $x^* = Tx^*$, any sequence obtained by recursively applying T, namely $x^{k+1} = Tx^k$, converges to a fixed point. Our main result (Theorem 3 below) states that our generalized proximal operator introduced in section 1 is BFNE with respect to a hybrid Bregman divergence mixing the squared Euclidean and the KL divergences.

[1] see e.g. wikipedia.org/Minimax_theorem or Theorem 7.1 in Cesa-Bianchi and Lugosi [2006]

Definition 2 (Euclidean+KL Bregman divergence)**.** Let the function f be defined for all $(x, q) \in \mathbb{R}^m \times \Delta_S$ as follows:

$$f(x, q) = \frac{1}{2}\|x\|^2 + h(q)$$

and its corresponding Bregman divergence:

$$D_f\left(\begin{pmatrix} x \\ q \end{pmatrix}, \begin{pmatrix} x' \\ q' \end{pmatrix}\right) = \frac{1}{2}\|x - x'\|^2 + D_{\mathrm{KL}}(q\|q').$$

Theorem 3 (BFNE)**.** *Let $\mathrm{prox}_{\lambda\ell}$ and f be as defined in Definitions 1 and 2 respectively. Then, $\mathrm{prox}_{\lambda\ell}$ is Bregman firmly nonexpansive with respect to f.*

Proof. For $q \in \Delta_S$, we have by the convexity of $x \mapsto q^\intercal \ell(x)$:

$$q^\intercal \ell(z) - q^\intercal \ell(x) \geq q^\intercal J_\ell(x)(z - x) \tag{2}$$

and, similarly, for any other $r \in \Delta_S$:

$$r^\intercal \ell(x) - r^\intercal \ell(z) \geq r^\intercal J_\ell(z)(x - z). \tag{3}$$

Then, by summing Eqs. 2 and 3 it holds:

$$\begin{aligned}(J_\ell(z)^\intercal r - J_\ell(x)^\intercal q)^\intercal (z - x) &\geq q^\intercal \ell(x) - q^\intercal \ell(z) + r^\intercal \ell(z) - r^\intercal \ell(x) \\ &\iff \left\langle \begin{pmatrix} z \\ r \end{pmatrix} - \begin{pmatrix} x \\ q \end{pmatrix}, \begin{pmatrix} J_\ell(z)^\intercal r \\ -\ell(z) \end{pmatrix} - \begin{pmatrix} J_\ell(x)^\intercal q \\ -\ell(x) \end{pmatrix} \right\rangle \geq 0. \end{aligned} \tag{4}$$

From the stationarity condition satisfied by the saddle point $(x', q') = \mathrm{prox}_\lambda(x, q)$ of the function $H_{x,q}$:

$$\nabla H_{x,q}(x', q') = 0 \Leftrightarrow \begin{cases} J_\ell(x')^\intercal q' + \frac{1}{\lambda}(x' - x) = 0 \\ \ell(x') - \frac{1}{\lambda}(\nabla h(q') - \nabla h(q)) = 0 \end{cases} \Leftrightarrow \begin{cases} x = x' + \lambda J_\ell(x')^\intercal q' \\ \nabla h(q) = \nabla h(q') - \lambda \ell(x'). \end{cases} \tag{5}$$

We are now ready to prove that $\mathrm{prox}_{\lambda\ell}$ is BFNE w.r.t. f. For $x, z \in \mathbb{R}^m$, $q, r \in \Delta_S$ and $(x', q') = \mathrm{prox}_{\lambda\ell}(x, q)$, $(z', r') = \mathrm{prox}_{\lambda\ell}(z, r)$

$$\begin{aligned}&\langle \mathrm{prox}_{\lambda\ell}(x, q) - \mathrm{prox}_{\lambda\ell}(z, r), \nabla f(x, q) - \nabla f(z, r)\rangle = \\ &\qquad \left\langle \begin{pmatrix} x' \\ q' \end{pmatrix} - \begin{pmatrix} z' \\ r' \end{pmatrix}, \begin{pmatrix} x \\ \nabla h(q) \end{pmatrix} - \begin{pmatrix} z \\ \nabla h(r) \end{pmatrix} \right\rangle \\ &\qquad = \left\langle \begin{pmatrix} x' - z' \\ q' - r' \end{pmatrix}, \begin{pmatrix} x' + \lambda J_\ell(x')^\intercal q' - z' - \lambda J_\ell(z')^\intercal r' \\ \nabla h(q') - \lambda \ell(x') - \nabla h(r') + \lambda \ell(z') \end{pmatrix} \right\rangle \\ &\qquad = \|x' - z'\|^2 + \langle q' - r', \nabla h(q') - \nabla h(r')\rangle \\ &\qquad + \lambda\langle x' - z', J_\ell(x')^\intercal q' - J_\ell(z')^\intercal r'\rangle + \lambda\langle q' - r', -\ell(x') + \ell(z')\rangle \\ &\geq \|x' - z'\|^2 + \langle q' - r', \nabla h(q') - \nabla h(r')\rangle = \left\langle \begin{pmatrix} x' \\ q' \end{pmatrix} - \begin{pmatrix} z' \\ r' \end{pmatrix}, \nabla f(x', q') - \nabla f(z', r') \right\rangle \end{aligned} \tag{6}$$

where the inequality comes from Eq. (4).

□

We highlight that Theorem 3 generalizes the fact that the classic proximal operator is firmly nonexpansive, since D_f reduces to the squared Euclidean Bregman divergence in the convex scenario $S = 1$. Moreover, the next result shows that our prox can also be written as a generalized resolvent. Indeed, we recall from Eckstein [1993]-Bauschke et al. [2003]-Borwein et al. [2011] that an f-resolvent is equal to $(\nabla f + \lambda A)^{-1} \circ \nabla f$ for some monotone operator A. This definition extends the classic notion of resolvent, namely $(I + \lambda A)^{-1}$ (which corresponds to the particular case $f = \frac{\|\cdot\|^2}{2}$), to a general Bregman divergence D_f.

Proposition 4 (f-resolvent)**.** *Consider the notations introduced in Definition 1.*

(i) The operator $A(x, q) = \begin{pmatrix} J_\ell(x)^\intercal q \\ -\ell(x) \end{pmatrix}$ *is monotone.*

(ii) Our prox operator is an f-resolvent:

$$prox_{\lambda\ell} = (\nabla f + \lambda A)^{-1} \circ \nabla f \ ,$$

with A from (i) and f from Definition 2.

Proof. (i) follows from the inequality in Eq. (4) while (ii) derives from the stationarity condition (Eq. 5) of the saddle point $(x', q') = \text{prox}_{\lambda\ell}(x, q)$ of the function $H_{x,q}$. □

PPA and fixed point Theorem 3 implies that the generalized proximal point algorithm $(x^{k+1}, q^{k+1}) = \text{prox}_{\lambda\ell}(x^k, q^k)$ converges to a fixed point (x^*, q^*) of the prox, if there exists any. Such a fixed point is characterized by:

$$(x^*, q^*) = \text{prox}_{\lambda\ell}(x^*, q^*) \Leftrightarrow \begin{cases} J_\ell(x^*)^\intercal q^* = 0 \\ q^*_s = \frac{q^*_s e^{-\lambda \ell_s(x^*)}}{\sum_{s'} q^*_{s'} e^{-\lambda \ell_{s'}(x^*)}} \quad (\forall 1 \leq s \leq S) \end{cases} \tag{7}$$

which means that the q^*-barygradient of ℓ at x^* is equal to zero and that for all $(s, t) \in \{1, \ldots, S\}^2$:

$$q^*_s \neq 0 \text{ and } q^*_t \neq 0 \Rightarrow \ell_s(x^*) = \ell_t(x^*). \tag{8}$$

3 The BGD algorithm

By analogy with the gradient descent update rule that approximates (for small λ) the classic prox, we propose in the following to derive from our generalized prox an algorithm that we call the *barygradient descent algorithm* (or BGD in short). To do so, we leverage the equation (1) by using the $\min_z \max_r$ (resp.

$\max_r \min_z$) characterization of the saddle point $(x', q') = \mathrm{prox}_{\lambda\ell}(x, q)$ to update x (resp. q).

From now on, let us assume that ℓ is twice continuously differentiable and that $\max_{1\leq s\leq S} \sup_z \rho(\nabla^2 \ell_s(z)) < \infty$ with $\rho(\nabla^2 \ell_s(z))$ the spectral radius of the Hessian of ℓ_s at z.

3.1 Updating x ("maximize then minimize")

In order to approximate x', we use the equality $H_{x,q}(x', q') = \min_z \max_r H_{x,q}(z, r)$ from Eq. (1). In other words, we first compute the maximization over the simplex for a given z, which boils down to a mirror descent update (see Beck and Teboulle [2003], Bubeck et al. [2015]). Indeed, the solution $r(z) = \arg\max_{r\in\Delta_S} H_{x,q}(z, r)$ satisfies:

$$\nabla h(r(z)) = \nabla h(q) + \lambda \ell(z) \quad \Longleftrightarrow \forall s, \quad r_s(z) = \frac{q_s e^{\lambda \ell_s(z)}}{\sum_{s'} q_{s'} e^{\lambda \ell_{s'}(z)}}. \tag{9}$$

Then by injecting $r(z)$ from Eq. (9) into $H_{x,q}$ we obtain:

$$H_{x,q}(z, r(z)) = \underbrace{\frac{1}{\lambda} \log\left(\sum_s q_s e^{\lambda \ell_s(z)}\right)}_{F_q(z)} + \frac{1}{2\lambda}\|z - x\|^2. \tag{10}$$

Observe that the left Bregman-Moreau envelope $q \mapsto F_q(z) = \max_{r\in\Delta_S} r^\intercal \ell(z) - \frac{1}{\lambda} D_{\mathrm{KL}}(r\|q)$ is concave (see Bauschke et al. [2018]) while $z \mapsto F_q(z)$ is convex as a supremum of convex functions. When replacing the convex function $F_q(\cdot)$ in Eq. (10) by its 1st order Taylor approximation, namely $\hat{F}_q(z) = F_q(x) + \nabla F_q(x)^\intercal (z - x)$, the minimizer of $\hat{F}_q(z) + (1/2\lambda)\|z - x\|^2$ is simply given by $\hat{z} = x - \lambda \nabla F_q(x)$ (GD update) with gradient

$$\boxed{\nabla F_q(x) = \frac{\sum_s q_s e^{\lambda \ell_s(x)} \nabla \ell_s(x)}{\sum_s q_s e^{\lambda \ell_s(x)}}}. \tag{11}$$

Note that $\nabla F_q(x)$ can be interpreted as the $\hat{q}$-barygradient of ℓ at x with $\hat{q}_s \propto q_s e^{\lambda \ell_s(x)}$.

3.2 Updating q ("minimize then maximize")

Symmetrically, let us now approximate q' by using the equality $H_{x,q}(x', q') = \max_r \min_z H_{x,q}(z, r)$ from Eq. (1) (i.e. first minimize w.r.t. z then maximize w.r.t. r). Given $r \in \Delta_S$, the solution $z(r) = \arg\min_{z\in\mathbb{R}^m} H_{x,q}(z, r)$ is a standard PPA update (for the convex function $z \mapsto r^\intercal \ell(z)$) given by

$$z(r) = (I + \lambda J_\ell^\intercal r)^{-1} x. \tag{12}$$

Then by plugging Eq. (12) into $H_{x,q}(z(r), r)$ we obtain the following concave maximization problem over the probability simplex:

$$\max_{r\in\Delta_S} \underbrace{r^\intercal \ell((I+\lambda J_\ell^\intercal r)^{-1}x) + \frac{\lambda}{2}\|J_\ell((I+\lambda J_\ell^\intercal r)^{-1}x)^\intercal r\|^2}_{E_x(r)} - \frac{1}{\lambda} D_{\mathrm{KL}}(r\|q). \quad (13)$$

Here again we propose to approximate the problem above by replacing the concave function E_x by its 1st order Taylor expansion and obtain a mirror ascent update rule, namely $\nabla h(\hat{r}) = \nabla h(q) + \lambda \nabla E_x(q)$.

Gradient of E_x. Let us denote $B(q) = I + \lambda J_\ell^\intercal q$. We have:

$$\begin{aligned} E_x(q) &= q^\intercal \ell(B(q)^{-1}x) + \frac{\lambda}{2}\|J_\ell(B(q)^{-1}x)^\intercal q\|^2 \\ &= \sum_s q_s \ell_s(B(q)^{-1}x) + \frac{\lambda}{2}\|\sum_s q_s \nabla \ell_s(B(q)^{-1}x)\|^2 . \quad (14) \end{aligned}$$

• First term: Denoting $y = B(q)^{-1}x$, we have for the derivative of the first term in Eq. (14):

$$\frac{\partial}{\partial q_t}\left(\sum_s q_s \ell_s(B(q)^{-1}x)\right) = \ell_t(y) + \sum_s q_s \frac{\partial \ell_s(y)}{\partial q_t}. \quad (15)$$

Now $\frac{\partial \ell_s(y)}{\partial q_t}$ involves differentiating y w.r.t. q_t. Using the formula for the derivative of an inverse operator, we have:

$$\frac{\partial}{\partial q_t} B(q)^{-1} = -B(q)^{-1} \underbrace{\frac{\partial B(q)}{\partial q_t}}_{\lambda \nabla \ell_t} B(q)^{-1} ,$$

which implies $\frac{\partial y}{\partial q_t} = -\lambda B(q)^{-1}\nabla \ell_t(y)$. Therefore

$$\frac{\partial \ell_s(y)}{\partial q_t} = \nabla \ell_s(y)^\intercal \frac{\partial y}{\partial q_t} = -\lambda \nabla \ell_s(y)^\intercal B(q)^{-1} \nabla \ell_t(y).$$

Hence, the gradient (w.r.t. q_t) of the first term is

$$\ell_t(y) - \lambda \sum_s q_s \nabla \ell_s(y)^\intercal B(q)^{-1} \nabla \ell_t(y). \quad (16)$$

• Second term: For the derivative of the second term in Eq. (14), denoting $Z = \sum_s q_s \nabla \ell_s(B(q)^{-1}x)$, the gradient of $\frac{\lambda}{2}\|Z\|^2$ with respect to q_t is: $\lambda Z^\intercal \frac{\partial Z}{\partial q_t}$. Now, compute

$$\frac{\partial Z}{\partial q_t} = \nabla \ell_t(y) + \sum_s q_s \frac{\partial}{\partial q_t} \nabla \ell_s(y). \quad (17)$$

Similarly to the first term, differentiating $\nabla \ell_s(y)$ w.r.t. q_t gives:

$$\frac{\partial}{\partial q_t}\nabla \ell_s(y) = -\lambda [\nabla_j^2 \ell_s(y)^\intercal B(q)^{-1} \nabla \ell_t(y)]_{1\le j\le m} = -\lambda \nabla^2 \ell_s(y)^\intercal B(q)^{-1} \nabla \ell_t(y),$$

where $\nabla^2 \ell_s(y)$ denotes the Hessian matrix of ℓ_s at y, and $\nabla^2_j \ell_s(y)$ its j-th column. Finally, the gradient of the second term is:

$$\lambda Z^\intercal \frac{\partial Z}{\partial q_t} = \lambda \left(\sum_s q_s \nabla \ell_s(y) \right)^\intercal \left(\nabla \ell_t(y) - \lambda \sum_s q_s \nabla^2 \ell_s(y)^\intercal B(q)^{-1} \nabla \ell_t(y) \right). \tag{18}$$

• Collecting all terms: By summing Eqs. 16-18 we conclude that $\forall t \in \{1, \dots, S\}$,

$$\boxed{\frac{\partial E_x}{\partial q_t}(q) = \ell_t(R(q)x) + \lambda^2 \sum_s q_s \nabla \ell_s(R(q)x)^\intercal \left(Y(q) - \sum_{s'} q_{s'} \nabla^2 \ell_{s'}(R(q)x)^\intercal R(q) \right) \nabla \ell_t(R(q)x)}, \tag{19}$$

with the resolvent $R(q) = B(q)^{-1}$ and $Y(q) = \frac{1}{\lambda}(I - B(q)^{-1})$ the Yosida approximation of the q-barygradient of ℓ. In practice $R(q)$ can be approximated by the Neumann series $I - \lambda J_\ell^\intercal q + \mathcal{O}_{\|\cdot\|_{\mathrm{op}}}(\lambda^2)$ for $\lambda < (\sup_z \rho(\sum_s q_s \nabla^2 \ell_s(z)))^{-1}$.

3.3 Barygradient descent

Given $(x^k, q^k) \in \mathbb{R}^m \times \Delta_S$, we define the next BGD iterate (x^{k+1}, q^{k+1}) as follows (by combining sections 3.1 and 3.2):

$$\boxed{\begin{cases} x^{k+1} = x^k - \lambda \nabla F_{q^k}(x^k) \\ \nabla h(q^{k+1}) = \nabla h(q^k) + \lambda \nabla E_{x^k}(q^k) \end{cases}}. \tag{BGD}$$

Acknowledgments

The author thanks Adil Salim and Massil Achab for helpful discussions on convex analysis.

References

Heinz H Bauschke and Patrick L Combettes. Convex analysis and monotone operator theory in hilbert spaces. 2011.

Heinz H Bauschke, Jonathan M Borwein, and Patrick L Combettes. Bregman monotone optimization algorithms. *SIAM Journal on control and optimization*, 42(2):596–636, 2003.

Heinz H Bauschke, Minh N Dao, and Scott B Lindstrom. Regularizing with bregman–moreau envelopes. *SIAM Journal on Optimization*, 28(4):3208–3228, 2018.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3): 167–175, 2003.

Jonathan M Borwein, Simeon Reich, and Shoham Sabach. A characterization of bregman firmly nonexpansive operators using a new monotonicity concept. *J. Nonlinear and Convex Analysis*, 12:161–183, 2011.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Myrana Brohé and Patricia Tossings. Perturbed proximal point algorithm with nonquadratic kernel. *Serdica Math. J*, 26:177–206, 2000.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Jonathan Eckstein. Nonlinear proximal point algorithms using bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.