A Fast Proximal Point Method for Computing Exact Wasserstein Distance – Appendix

A More Analysis on IPOT

A.1 Convergence w.r.t. L

As mentioned in Section 6.1, we provide the test result of 64D Gaussian distributed data here. We choose the computed Wasserstein distance $\langle \mathbf{\Gamma}, \boldsymbol{C} \rangle$ as the indicator of convergence, because while the optimal transport plan might not be unique, the computed Wasserstein distance at convergence must be unique and minimized to ground truth. We use the empirical distribution as input distributions, i.e.,

$$\begin{aligned} &W(\{x_i\}, \{g_\theta(z_j)\}) = \min_{\mathbf{\Gamma}} \langle \boldsymbol{C}(\theta), \mathbf{\Gamma} \rangle \\ &\text{s.t. } \mathbf{\Gamma}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n, \mathbf{\Gamma}^T\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n. \end{aligned} \tag{14}$$

As shown in Figure 8, the convergence rate is also linear. For comparison, we also provide the convergence path of Sinkhorn iteration. The result cannot converge to ground truth because the method is essentially regularized.

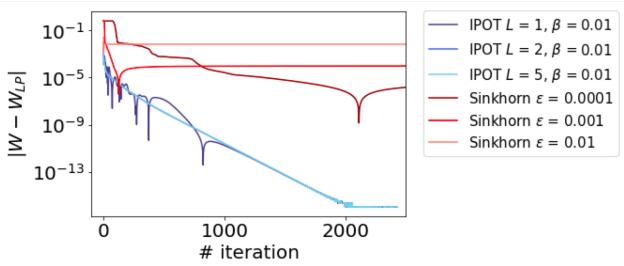


Figure 8: The plot of differences in computed Wasserstein distances w.r.t. number of iterations for 64D Gaussian distributed data. Here, W are the Wasserstein distance computed at current iteration. W_{LP} is computed by simplex method, and is used as ground truth. The test adopts $c(x, y) = ||x - y||_2$. Due to random data is used, the number of iteration that the algorithm reaches 10^{-17} varies from 1000 to around 5000 according to our tests.

Remark. When we are talking about amount of regularization, usually we are referring to the magnitude of ϵ for Sinkhorn, or the equivalent magnitude of ϵ computed from remark in Section 3 for IPOT method. However, the amount of regularization in a loss function should be quantified by $\epsilon/||\boldsymbol{C}||$, instead of ϵ alone. That is why in this paper, different magnitude of ϵ is used for different application.

A.2 How IPOT Avoids Instability

Heuristically, if Sinkhorn does not underflow, with enough iteration, the result of IPOT is approximately the same as Sinkhorn with $\epsilon^{(t)} = \beta/t$. The difference lies in IPOT is a principled way to avoid underflow and can converge to arbitrarily small regularization, while Sinkhorn always causes numerical difficulty when $\epsilon \to 0$, even with scheduled decreasing ϵ like [6]. More specifically, in IPOT, we can factor $\Gamma = \text{diag}(\boldsymbol{u}_1)\boldsymbol{G}^t\text{diag}(\boldsymbol{u}_2)$, where $(\cdot)^t$ is element-wise exponent operation, and $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are two scaling vectors. So we have $\epsilon^{(t)} = \beta/t$. As t goes infinity, all entries of $\boldsymbol{G}^t$ would underflow if we use Sinkhorn with $\epsilon^{(t)} = \beta/t$. But we know $\mathbf{\Gamma}^*$ is neither all zeros nor contains infinity. So instead of computing $\boldsymbol{G}^t$, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ directly, we use $\mathbf{\Gamma}^t$ to record the multiplication of $\boldsymbol{G}^t$ with part of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ in each step, so the entries of $\mathbf{\Gamma}^t$ will not over/underflow. The explicit computation of $\boldsymbol{G}^t$ is not needed.

Therefore, by tuning β and iteration number, we can achieve the result of arbitrary amount of regularization with IPOT.

B Learning Generative Models

In this section, we show the derivation for the learning algorithm, and more tests result.

For simplicity, we assume $|\{x_i\}| = |\{z_j\}| = n$. Given a dataset $\{x_i\}$ and some noise $\{z_j\}$ [4, 16], our goal is to find a

parameterized function $g_\theta(\cdot)$ that minimize $W(\{x_i\}, \{g_\theta(z_j)\})$,

$$\begin{aligned} W(\{x_i\}, \{g_\theta(z_j)\}) = & \min_{\mathbf{\Gamma}} \langle \boldsymbol{C}(\theta), \mathbf{\Gamma} \rangle \\ & \text{s.t. } \mathbf{\Gamma} \mathbf{1}_n = \frac{1}{n} \mathbf{1}_n, \mathbf{\Gamma}^T \mathbf{1}_n = \frac{1}{n} \mathbf{1}_n, \end{aligned} \tag{15}$$

where $\boldsymbol{C}(\theta) = [c(x_i, g_\theta(z_j))]$.Usually, g_θ is parameterized by a neural network with parameter θ, and the minimization over θ is done by stochastic gradient descent.

In particular, given current estimation θ, we can obtain optimum $\mathbf{\Gamma}^*$ by IPOT, and compute the Wasserstein distance by $\langle \boldsymbol{C}(\theta), \mathbf{\Gamma}^* \rangle$ accordingly. Then, we can further update θ by the gradient of current Wasserstein distance. There are two ways to solve the gradient: One is auto-diff based method such as [15], the other is based on the envelope theorem [1]. Different from the auto-diff based methods, the back-propagation based on envelope theorem does not go into proximal point iterations because the derivative over $\mathbf{\Gamma}^*$ is not needed, which accelerates the learning process greatly. This also has significant implications numerically because the derivative of a computed quantity tends to amplify the error. Therefore, we adopt envelope based method.

Theorem B.1. ***Envelope theorem.*** *Let $f(x, \theta)$ and $l(x)$ be real-valued continuously differentiable functions, where $x \in \mathbb{R}^n$ are choice variables and $\theta \in \mathbb{R}^m$ are parameters. Denote x^* to be the optimal solution of f with constraint $l = 0$ and fixed θ, i.e.*

$$x^* = \arg\min_x f(x, \theta) \quad s.t. \quad l(x) = 0.$$

Then, assume that V is continuously differentiable function defined as $V(\theta) \equiv f(x^*(\theta), \theta)$, the derivative of V over parameters is

$$\frac{\partial V(\theta)}{\partial \theta} = \frac{\partial f}{\partial \theta}.$$

In our case, because $\mathbf{\Gamma}^*$ is the minimization of $\langle \mathbf{\Gamma}, \boldsymbol{C}(\theta) \rangle$ with constraints, we have

$$\begin{aligned} & \frac{\partial W(\{x_i\}, \{g_\theta(z_j)\})}{\partial \theta} = \frac{\partial \langle \mathbf{\Gamma}^*, \boldsymbol{C}(\theta) \rangle}{\partial \theta} \\ & = \langle \mathbf{\Gamma}^*, \frac{\partial \boldsymbol{C}(\theta)}{\partial \theta} \rangle = \langle \mathbf{\Gamma}^*, 2(g_\theta(z_j) - x_i) \frac{\partial g_\theta(z_i)}{\partial \theta} \rangle, \end{aligned}$$

where we assume $C_{ij}(\theta) = \|x_i - g_\theta(z_j)\|_2^2$, but the algorithm can also adopt other metrics. The derivation is in supplementary materials. The flowchart is shown in Figure 9, and the algorithm is shown in Algorithm 3.

Note Sinkhorn distance is defined as $S(\{x_i\}, \{g_\theta(z_j)\}) = \langle \boldsymbol{C}(\theta), \mathbf{\Gamma}^* \rangle$, where $\mathbf{\Gamma}^* = \arg\min_{\mathbf{\Gamma} \in \Sigma(\mathbf{1}/n, \mathbf{1}/n)} \langle \boldsymbol{C}(\theta), \mathbf{\Gamma} \rangle + \epsilon h(\mathbf{\Gamma})$. If Sinkhorn distance is used in learning generative models, envelope theorem cannot be used because the loss function for optimizing θ and $\mathbf{\Gamma}$ is not the same.

In the tests, we observe the method in [15] suffers from shrinkage problem, i.e. the generated distribution tends to shrink towards the target mean. The recovery of target distribution is sensitive to the weight of regularization term ϵ. Only relatively small ϵ can lead to a reasonable generated distribution.

Algorithm 3 Learning generative networks

Input: real data $\{x_i\}$, initialized generator g_θ
while not converged **do**
  Sample a batch of real data $\{x_i\}_{i=1}^n$
  Sample a batch of noise data $\{z_j\}_{i=1}^n \sim q$
  $C_{ij} := c(x_i, g_\theta(z_j)) := ||x_i - g_\theta(z_j)||_2^2$
  $\mathbf{\Gamma}$ = IPOT($\frac{1}{n}\mathbf{1}_n, \frac{1}{n}\mathbf{1}_n, \boldsymbol{C}$)
  Update θ with $\langle \mathbf{\Gamma}, [2(x_i - g_\theta(z_j)) \frac{\partial g_\theta(z_j)}{\partial \theta}] \rangle$
end while

B.1 Synthetic Test

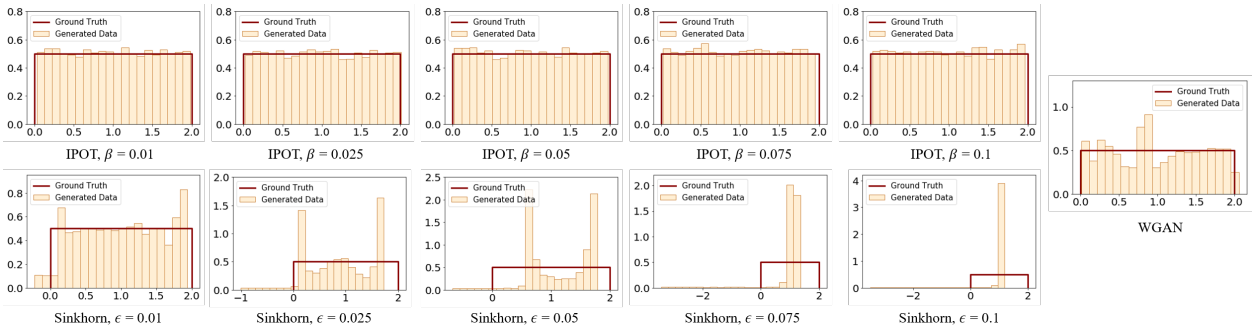


Figure 10: The sequences of learning result of IPOT, Sinkhorn. In each figure, the orange histogram is the histogram of generated data, while the red line represents the PDF of the ground truth of target distribution.

In section 5.1, we show the learning result of Sinkhorn and IPOT in 2D case. In Figure 10 we show sequences of results for a 1D-1D generator, respectively. The upper sequence is IPOT with $\beta = 0.01, 0.025, 0.05, 0.075, 0.1$. The results barely change w.r.t. β. The lower sequence is the corresponding Sinkhorn results. The results shrink to the mean of target data, as expected. Also, we observe the learned distribution tends to have a tail that is not in the range of target data (also in 2D result, we do not include that part for a better view). It might be because the range of support that has a small probability has very small gradient when updated. Once the distribution is initialized to have a tail with small probability, it can hardly be updated. But this theory cannot explain why larger ϵ corresponds to longer tails. The tails can be on the left or right. We pick the ones on the left for easier comparison.

B.2 MNIST Test

The same shrinkage can be observed in MNIST data as well. See figure 11. While $\epsilon = 0.1$ covers most shapes of the numbers, $\epsilon = 1$ only covers a fraction, and $\epsilon = 10$ seems to cover only the mean of images.

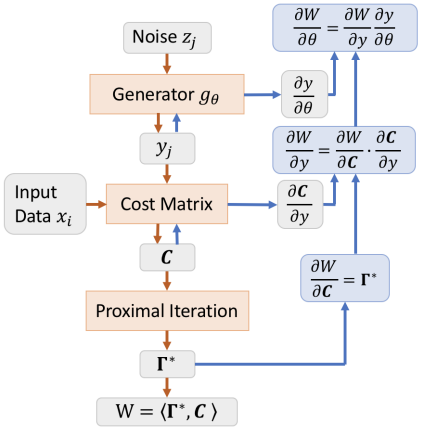


Figure 9: The architecture of the learning model using Envelope theorem in detail. According to Envelope theorem, we do not need to compute $\frac{\partial W}{\partial \Gamma^*}$, so we do not need to back-propagate into the iteration.

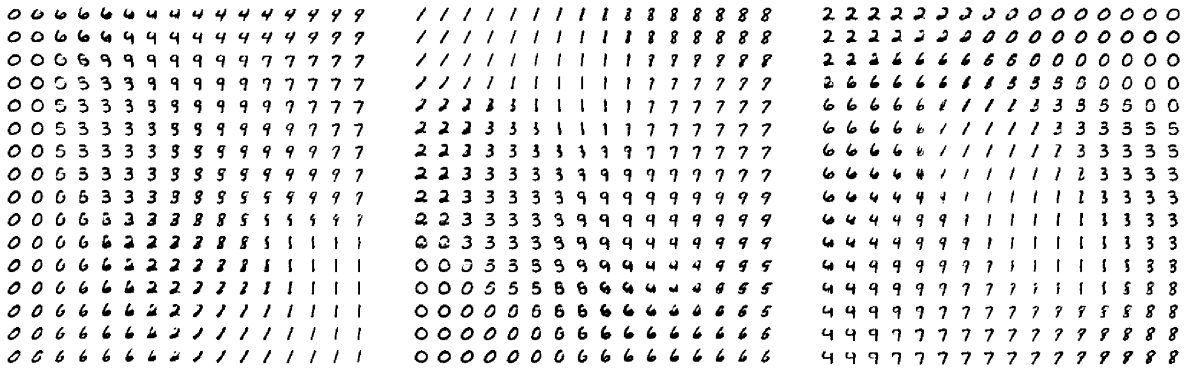
C Color Transferring

Optimal transport is directly applicable to many applications, such as color transferring and histogram calibration. We will show the result of color transferring and why accurate transport plan is superior to entropically regularized ones.

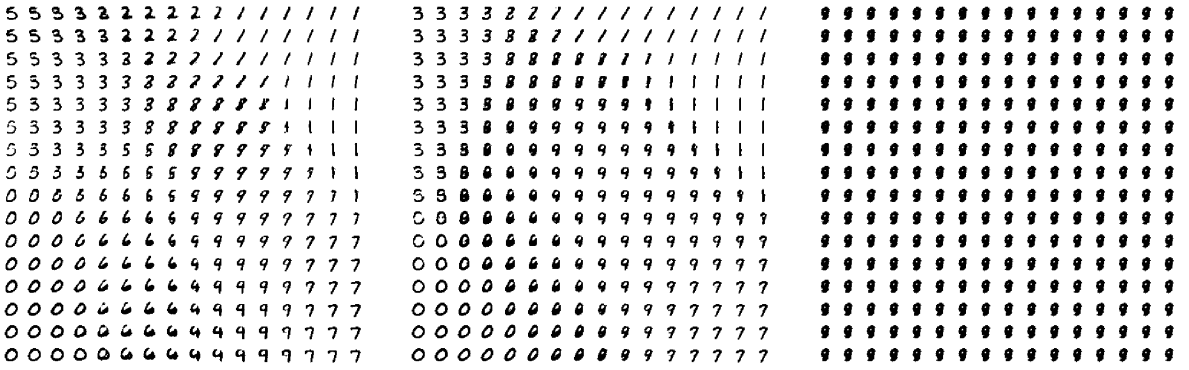
The goal of color transferring is to transfer the tonality of a target image into a source image. This is usually done by imposing the histogram of the color palette of one image to another image. Since Reinhard et al. [26], many methods [24, 42] are developed to do so by learning the transformation between the two histograms. Experiments in [30] have shown that transformation based on optimal transport map outperforms state-of-the-art techniques for challenging images.

Same as other prime-form Wasserstein distance solvers [22, 8], the proximal point method provide a transport plan. By definition, the plan is a transport from the source distribution to a target one with minimum cost. Therefore it provides a way to transform a histogram to another.

One example is shown in figure 12. We use three different maps to transform the RGB channels, respectively. For each channel, there are at most 256 bins. Therefore, using three channels separately is more efficient than treating the colors



(a) IPOT $\beta = 0.1$ (b) IPOT $\beta = 1$ (c) IPOT $\beta = 10$



(d) Sinkhorn $\epsilon = 0.1$ (e) Sinkhorn $\epsilon = 1$ (f) Sinkhorn $\epsilon = 10$

Figure 11: Plots of MNIST learning result under comparable resources with different ϵ. They both use batch size=200, number of hidden layer=1, number of nodes of hidden layer=500, number of iteration=500, learning rate = 1e-4. Note that despite we show result of $\epsilon = 0.1$ here, the algorithm does not run stably. It would sometimes fail due to numerical issue.

(a) Original image (b) Color image (c) Simplex

(d) Sinkhorn ϵ=10^{-2} (e) Sinkhorn ϵ=10^{-3} (f) IPOT

Figure 12: An example of color transferring. The right upper corner of each generated image shows the zoom-in of the color detail of the mouth corner.

as 3D data. Figure 12 shows proximal point method can produce identical result as linear programming at convergence, while the results produced by Sinkhorn method differ w.r.t. ϵ.

D General Bregman Proximal Point Algorithm

In the main body of the paper, we discussed the proximal point algorithm with specific Bregman distance, which is generated through the traditional entropy function. In this section, we generalize our results by proving the effectiveness of proximal point algorithm with general Bregman distance. Bregman distance is applied to measure the discrepancy between different matrices which turns out to be one of the key ideas in regularized optimal transport problems. Its special structure also give rise to proximal-type algorithms and projectors in solving optimization problems.

D.1 Basic Algorithm Framework and Preliminaries

The fundamental iterative scheme of general Bregman proximal point algorithm can be denoted as

$$x^{(t+1)} = \arg\min_{x \in X} \left\{ f(x) + \beta^{(t)} D_h(x, x^{(t)}) \right\}, \tag{16}$$

where $t \in \mathbb{N}$ is the index of iteration, and $D_h(x, x^{(t)})$ denotes a general Bregman distance between x and $x^{(t)}$ based on a Legendre function h (The definition is presented in the following). In the main body of the paper, h is specialized as the classical entropy function and as follows the related Bregman distance reduces to the generalized KL divergence. Furthermore, the Sinkhorn-Knopp projection can be introduced to compute each iterative subproblem. In the following, we present some fundamental definitions and lemmas.

Definition D.1. *Legendre function: Let $h : X \to (-\infty, \infty]$ be a lsc proper convex function. It is called*

1. Essentially smooth*: if h is differentiable on int dom h, with moreover $\|\nabla h(x^{(t)})\| \to \infty$ for every sequence $\{x^{(t)}\} \subset$ int dom h converging to a boundary point of dom h as $t \to +\infty$;*

2. Legendre type*: if h is essentially smooth and strictly convex on int dom h.*

Definition D.2. *Bregman distance: any given Legendre function h,*

$$D_h(x,y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle, \quad \forall x \in \textit{dom}\, h, \forall y \in \textit{int dom}\, h, \tag{17}$$

where D_h is strictly convex with respect to its first argument. Moreover, $D_h(x,y) \geq 0$ for all $(x,y) \in$ dom $h \times$ int dom h, and it is equal to zero if and only if $x = y$. However, D_h is in general asymmetric, i.e., $D_h(x,y) \neq D_h(y,x)$.

Definition D.3. *Symmetry Coefficient: Given a Legendre function $h : X \to (-\infty, \infty]$, its symmetry coefficient is defined by*

$$\alpha(h) = \inf \left\{ \frac{D_h(x,y)}{D_h(y,x)} \;\middle|\; (x,y) \in \textit{int dom}\, h \times \textit{int dom}\, h,\ x \neq y \right\} \in [0,1]. \tag{18}$$

Lemma D.4. *Given $h : X \to (-\infty, +\infty]$, D_h is general Bregman distance, and x, y, $z \in X$ such that $h(x)$, $h(y)$, $h(z)$ are finite and h is differentiable at y and z,*

$$D_h(x,z) - D_h(x,y) - D_h(y,z) = \left\langle \nabla h(y) - \nabla h(z), x - y \right\rangle \tag{19}$$

Proof. The proof is straightforward as one can easily verify it by simply subtracting $D_h(y,z)$ and $D_h(x,y)$ from $D_h(x,z)$. □

D.2 Theorem 5.1 and Theorem 5.2

In this section, we first establish the convergence of Bregman proximal point algorithm, *i.e.*, **Theorem 5.1**, while our analysis is based on ([11, 12, 39]). Further, we establish the convergence of inexact version Bregman proximal point algorithm, *i.e.*, **Theorem 5.2**, in which the subproblem in each iteration is computed inexactly within finite number of sub-iterations.

Note that here for simplicity we provide proof of $d(\mathbf{\Gamma}, \mathbf{\Gamma}^{(t)}) = D_h(\mathbf{\Gamma}, \mathbf{\Gamma}^{(t)})$, i.e., the IPOT case. We can analogously prove it for $d(\{\mathbf{\Gamma}_k\}, \{\mathbf{\Gamma}_k^{(t)}\}) = \sum_{k=1}^K \lambda_k D_h(\mathbf{\Gamma}_k, \mathbf{\Gamma}_k^{(t)})$, i.e., the IPOT-WB case, with very similar proof. This is because the latter is essentially just the weighted version of the former.

Before proving both theorems, we propose several fundamental lemmas. The first Lemma is the fundamental descent lemma, which is popularly used to analysis the convergence result of first-order methods.

Lemma D.5. *(Descent Lemma) Consider a closed proper convex function $f : X \to (-\infty, \infty]$ and for any $x \in X$ and $\beta^{(t)} > 0$, we have:*

$$f(x^{(t+1)}) \leq f(x) + \beta^{(t)} \left[D_h(x, x^{(t)}) - D_h(x, x^{(t+1)}) - D_h(x^{(t+1)}, x^{(t)}) \right], \quad \forall x \in X. \tag{20}$$

Proof. The optimality condition of (16) can be written as

$$\left(x - x^{(t+1)} \right)^T \left[\nabla f(x^{(t+1)}) + \beta^{(t)} \left(\nabla h(x^{(t+1)}) - \nabla h(x^{(t)}) \right) \right] \geq 0, \quad \forall x \in X.$$

Then with the convexity of f, we obtain

$$f(x) - f(x^{(t+1)}) + \beta^{(t)} \left(x - x^{(t+1)} \right)^T \left(\nabla h(x^{(t+1)}) - \nabla h(x^{(t)}) \right) \geq 0. \tag{21}$$

With (19) it follows that

$$\left(x - x^{(t+1)} \right)^T \left(\nabla h(x^{(t+1)}) - \nabla h(x^{(t)}) \right) = D_h(x, x^{(t)}) - D_h(x, x^{(t+1)}) - D_h(x^{(t+1)}, x^{(t)}).$$

Substitute the above equation into (21), we have

$$f(x^{(t+1)}) \leq f(x) + \beta^{(t)} \left[D_h(x, x^{(t)}) - D_h(x, x^{(t+1)}) - D_h(x^{(t+1)}, x^{(t)}) \right], \quad \forall x \in X.$$

□

Next, we prove the convergence result in **Theorem 5.1**.

Theorem 5.1 *Let $\{x^{(t)}\}$ be the sequence generated by the general Bregman proximal point algorithm with iteration (16) where f is assumed to be continuous and convex. Further assume that $f^* = \min f(x) > -\infty$. Then we have that $\{f(x^{(t)})\}$ is non-increasing, and $f(x^{(t)}) \to f^*$. Further assume there exists η, s.t.*

$$f^* + \eta d(x) \le f(x), \quad \forall x \in X, \tag{22}$$

The algorithm has linear convergence.

Proof. 1. First, we prove the sufficient decrease property:

$$f(x^{(t+1)}) \le f(x^{(t)}) - \beta^{(t)}(1+\alpha(h))D_h(x^{(t+1)}, x^{(t)}). \tag{23}$$

Let $x = x^{(t)}$ in (20), we obtain

$$\begin{aligned} f(x^{(t+1)}) &\le f(x^{(t)}) - \beta^{(t)}\left[D_h(x^{(t)}, x^{(t+1)}) + D_h(x^{(t+1)}, x^{(t)})\right] \\ &\le f(x^{(t)}) - \beta^{(t)}(1+\alpha(h))D_h(x^{(t+1)}, x^{(t)}). \end{aligned}$$

With the sufficient decrease property, it is obvious that $\{f(x^{(t)})\}$ is non-decreasing.

2. Summing (23) from $i = 0$ to $i = t-1$ and for simplicity assuming $\beta^{(t)} = \beta$, we have

$$\sum_{i=0}^{k-1}\left[\frac{1}{\beta^{(t)}}\left(f(x^{(i+1)}) - f(x^{(i)})\right)\right] \le -\left[1+\alpha(h)\right]\sum_{i=0}^{k-1} D_h(x^{(i+1)}, x^{(i)})$$

$$\Rightarrow \qquad \sum_{i=0}^{\infty} D_h(x^{(i+1)}, x^{(i)}) < \frac{1}{\beta(1+\alpha(h))} f(x^{(0)}) < \infty,$$

which indicates that $D_h(x^{(i+1)}, x^{(i)}) \to 0$. Then summing (20) from $i = 0$ to $i = t-1$, we have

$$k\left(f(x^{(t)}) - f(x)\right) \le \sum_{i=0}^{t-1}\left(f(x^{(i+1)}) - f(x)\right) \le \beta D_h(x, x^{(0)}) < \infty, \quad \forall x \in X.$$

Let $t \to \infty$, we have $\lim_{t\to\infty} f(x^{(t)}) \le f(x)$ for every x, as a result we have $\lim_{k\to\infty} f(x^{(t)}) = f^*$.

3. Finally, we prove the convergence rate is linear. Assume $x^* = \arg\min_x f(x)$ is the unique optimal solution. Denote $d(x) = D_h(x^*, x)$. Let also $\beta^{(t)} = \beta$, we will prove

$$\frac{d(x^{(t+1)})}{d(x^{(t)})} \le \frac{1}{1+\frac{\eta}{\beta}} \tag{24}$$

Replace x with x^* in inequality (20), we have

$$f(x^{(t+1)}) \le f^* + \beta\left[d(x^{(t)}) - d(x^{(t+1)}) - D_h(x^{(t+1)}, x^{(t)})\right]. \tag{25}$$

Using assumption 22, we have

$$f^* + \eta d(x^{(t+1)}) \le f(x^{(t+1)}) \tag{26}$$

Sum 25 and 26 up, we have

$$\begin{aligned} \frac{\eta}{\beta} d(x^{(t+1)}) &\le d(x^{(t)}) - d(x^{(t+1)}) - D_h(x^{(t+1)}, x^{(t)}) \\ &\le d(x^{(t)}) - d(x^{(t+1)}) \end{aligned}$$

Therefore,

$$\frac{d(x^{(t+1)})}{d(x^{(t)})} \le \frac{1}{1+\frac{\eta}{\beta}}$$

Therefore, we have a linear convergence in Bregman distance sense.

□

Assumption (22) does not always hold when f is linear. In our specific case, x is bounded in $[0,1]^{m\times n}$. More rigorously, we can prove the following lemma.

Lemma D.6. *Assume $\mathcal{X}$ is a bounded polyhedron, x^* is unique, $d(x)$ is an arbitrary nonnegative convex function with $d(x*)=0$. If f is linear, then there exist η, s.t. $f^*+\eta d(x)\le f(x)$.*

Proof. Since $\mathcal{X}$ is a bounded polyhedron, any $x\in\mathcal{X}$ can be expressed as $x=\sum_{i=0}^{n}\lambda_i e_i$, where e_i is the vertices of $\mathcal{X}$, n is finite, and $\sum\lambda_i=1$. Also f is linear, so $f(x)=\sum_{i=0}^{n}\lambda_i f(e_i)$

Since f is linear, $\mathcal{X}$ is polyhedral and x^* is unique, x^* is a vertex of X. Denote $e_0=x^*$.

Denote $\delta=\min_{i>0}f(e_i)-f^*$, then $\delta>0$, or else x^* is not unique. Denote $d_{max}=\max_{i>0}d(e_i)$. Take

$$\eta=\delta/d_{max},$$

we have

$$\begin{aligned}
f^*+\eta d(x) &= f^*+\eta d(\sum_{i=0}^{n}\lambda_i e_i)\\
\text{(Jensen's Inequality)} &\le f^*+\eta\sum_{i=0}^{n}\lambda_i d(e_i)\\
(d(e_0)=0) \quad &\le f^*+(1-\lambda_0)\eta d_{max}\\
&= f^*+(1-\lambda_0)\delta\\
&= \sum_{i=1}^{n}\lambda_i(f^*+\delta)+\lambda_0 f^*\\
&\le \sum_{i=0}^{n}\lambda_i f(e_i)\\
&= f(x).
\end{aligned}$$

□

For more general cases, if x^* is not unique, we can divide the vertices as optimal vertices and the rest vertices, instead of e_0 and the rest as above, the conclusion can be proved analogously. Furthermore, if $\mathcal{X}$ is not a polyhedron, as long as $\mathcal{X}$ is bounded, we can always prove the conclusion in a polyhedron $\mathcal{A}$ s.t. $\mathcal{X}\in\mathcal{A}$ and $x*$ is also the optimal solution of $\min_{x\in\mathcal{A}}f(x)$. Proof of more general cases can be found in [27] (This paper points out some fairly strong continuity properties that polyhedral multifunctions satisfy).

Inequality (24) shows how the convergence rate is linked to β. This is the reason we claim in Section 4.1 that a smaller β would lead to quicker convergence in exact case.

From above, we showed that the general Bregman proximal point algorithm with constant step size can guarantee convergence to the optimal solution f^*, and has linear convergence rate with some assumptions. Further, we prove the convergence result for the general Bregman proximal point algorithm with inexact scheme in **Theorem 5.2**.

Theorem 5.2 *Let $\{x^{(t)}\}$ be the sequence generated by the general Bregman proximal point algorithm with inexact scheme (i.e., finite number of inner iterations are employed). Define an error sequence $\{e^{(t)}\}$ where*

$$e^{(t+1)}\in\beta^{(t)}\left[\nabla f(x^{(t+1)})+\partial\iota_X(x^{(t+1)})\right]+\left[\nabla h(x^{(t+1)})-\nabla h(x^{(t)})\right], \tag{27}$$

where ι_X is the indicator function of set X. If the sequence $\{e^{(t)}\}$ satisfies $\sum_{k=1}^{\infty}\|e^{(t)}\|<\infty$ and $\sum_{k=1}^{\infty}\langle e^{(t)},x^{(t)}\rangle$ exists and is finite, then $\{x^{(t)}\}$ converges to x^∞ with $f(x^\infty)=f^*$. If the sequence $\{e^{(t)}\}$ satisfies that exist $\rho\in(0,1)$ such that$\|e^{(t)}\|\le\rho^t$, $\langle e^{(t)},x^{(t)}\rangle\le\rho^t$ and with assumption (22), then $\{x^{(t)}\}$ converges linearly.

Remark: If exact minimization is guaranteed in each iteration, the sequence $\{x^{(t)}\}$ will satisfy that

$$0 \in \beta^{(t)} \left[\nabla f(x^{(t+1)}) + \partial \iota_X(x^{(t+1)})\right] + \frac{1}{\beta^{(t)}} \left[\nabla h(x^{(t+1)}) - \nabla h(x^{(t)})\right].$$

As a result, with enough inner iteration, the guaranteed $e^{(t)}$ will goes to zero.

Proof. This theorem is extended from [12, Theorem 1], and we propose a brief proof here. The proof contains the following four steps:

1. We have for all $k \geq 0$, through the three point lemma

   $$D_h(x, x^{(t+1)}) = D_h(x, x^{(t)}) - D_h(x^{(t+1)}, x^{(t)}) - \langle \nabla h(x^{(t)}) - \nabla h(x^{(t+1)}), x^{(t+1)} - x \rangle,$$

   which indicates

   $$D_h(x, x^{(t+1)}) = D_h(x, x^{(t)}) - D_h(x^{(t+1)}, x^{(t)}) - \langle \nabla h(x^{(t)}) - \nabla h(x^{(t+1)}) + e^{(t+1)}, x^{(t+1)} - x \rangle + \langle e^{(t+1)}, x^{(t+1)} - x \rangle.$$

   Since $\frac{1}{\beta^{(t)}} \left[e^{(t+1)} + \nabla h(x^{(t)}) - \nabla h(x^{(t+1)})\right] \in \nabla f(x^{(t+1)}) + \partial \iota_X(x^{(t+1)})$ and $0 \in \nabla f(x^*) + \partial \iota_X(x^*)$ if x^* be the optimal solution, we have

   $$\begin{aligned} & \langle \nabla h(x^{(t)}) - \nabla h(x^{(t+1)}) + e^{(t+1)}, x^{(t+1)} - x^* \rangle \\ = \quad & \beta^{(t)} \langle \left[\frac{1}{\beta^{(t)}} \left(\nabla h(x^{(t)}) - \nabla h(x^{(t+1)}) + e^{(t+1)}\right)\right] - 0, x^{(t+1)} - x^* \rangle \geq 0, \end{aligned}$$

   because $\nabla f + \partial \iota_X$ is monotone ($f + \iota_X$ is convex). Further we have

   $$D_h(x^*, x^{(t+1)}) \leq D_h(x^*, x^{(t)}) - D_h(x^{(t+1)}, x^{(t)}) + \langle e^{(t+1)}, x^{(t+1)} - x^* \rangle.$$

2. Summing the above inequality from $i = 0$ to $i = t - 1$, we have

   $$D_h(x^*, x^{(t)}) \leq D_h(x^*, x^{(0)}) - \sum_{i=0}^{t-1} D_h(x^{(i+1)}, x^{(i)}) + \sum_{i=0}^{t-1} \langle e^{(i+1)}, x^{(i+1)} - x^* \rangle.$$

   Since $\sum_{t=1}^{\infty} \|e^{(t)}\| < \infty$ and $\sum_{t=1}^{\infty} \langle e^{(t)}, x^{(t)} \rangle$ exists and is finite, we guarantee that

   $$\bar{E}(x^*) = \sup_{t \geq 0} \left\{ \sum_{i=0}^{t-1} \langle e^{(i+1)}, x^{(i+1)} - x^* \rangle \right\} < \infty.$$

   Together with $D_h(x^{(i+1)}, x^{(i)}) > 0$, we have

   $$D_h(x^*, x^{(t)}) \leq D_h(x^*, x^{(0)}) + \bar{E}(x^*) < \infty,$$

   which indicates

   $$0 \leq \sum_{i=0}^{\infty} D_h(x^{(i+1)}, x^{(i)}) < D_h(x^*, x^{(0)}) + \bar{E}(x^*) < \infty,$$

   and hence $D_h(x^{(i+1)}, x^{(i)}) \to 0$.

3. Based on the above two items, we know that the sequence $\{x^{(t)}\}$ must be bounded and has at least one limit point x^∞. The most delicate part of the proof is to establish that $0 \in \nabla f(x^\infty) + \partial \iota_X(x^\infty)$. Let $T = \nabla f + \partial \iota_X$, then T denotes the subdifferential mapping of a closed proper convex function $f + \iota_X$ (f is a linear function and X is a closed convex set). Let $\{t_j\}$ be the sub-sequence such that $x^{t_j} \to x^\infty$. Because $x^{t_j} \in X$ and X is a closed convex set, we know $x^\infty \in X$. We know that $D_h(x^*, x^{(t+1)}) \leq D_h(x^*, x^{(t)}) + \langle e^{(t+1)}, x^{(t+1)} - x^* \rangle$

and $\sum_{k=0}^{\infty}\langle e^{(t+1)}, x^{(t+1)} - x^*\rangle$ exists and is finite. From [23, Section 2.2], we guarantee that $\{D_h(x^*, x^{(t)})\}$ converges to $0 \leq d(x^*) < \infty$. Define $y^{(t+1)} := \lambda_k \left(\nabla h(x^{(t)}) - \nabla h(x^{(t+1)}) + e^{(t+1)}\right)$, we have

$$\lambda_k\langle y^{(t+1)}, x^{(t+1)} - x^*\rangle = D_h(x^*, x^{(t)}) - D_h(x^*, x^{(t+1)}) - D_h(x^{(t+1)}, x^{(t)}) + \langle e^{(t+1)}, x^{(t+1)} - x^*\rangle.$$

By taking the limit of both sides and $\lambda_k = \lambda > 0$, we obtain that

$$\langle y^{(t+1)}, x^{(t+1)} - x^*\rangle \to 0.$$

For the reason that y^{k_j+1} is a subgradient of $f + \iota_X$ at x^{k_j+1}, we have

$$f(x^*) \geq f(x^{k_j+1}) + \langle y^{k_j+1}, x^* - x^{k_j+1}\rangle, \quad x^* \in X, x^{k_j+1} \in X.$$

Further let $j \to \infty$ and using f is lower semicontinuous, $\langle y^{(t+1)}, x^{(t+1)} - x^*\rangle \to 0$, we obtain

$$f(x^*) \geq f(x^\infty), \quad x^\infty \in X$$

which implies that $0 \in \nabla f(x^\infty) + \iota_X(x^\infty)$.

4. Recall the inexact scheme (27), we can equivalently guarantee that

$$(x - x^{(t+1)})^T \left\{\beta^{(t)}\nabla f(x^{(t+1)}) + \left[\nabla h(x^{(t+1)}) - \nabla h(x^{(t)})\right] - e^{(t+1)}\right\} \geq 0, \ \forall x \in X.$$

Together the convexity of f and the three point lemma, we obtain

$$f(x^{(t+1)}) \leq f(x) + \frac{1}{\beta^{(t)}}\left[D_h(x, x^{(t)}) - D_h(x, x^{(t+1)}) - D_h(x^{(t+1)}, x^{(t)}) - (x - x^{(t+1)})^T e^{(t+1)}\right].$$

Let $x = x^*$ in the above inequality and recall the assumption (22), *i.e.*,

$$f(x) - f(x^*) \geq \eta d(x),$$

we have with $\beta^{(t)} = \beta$

$$\begin{aligned}\eta d(x^{(t+1)}) &\leq \frac{1}{\beta}\left[d(x^{(t)}) - d(x^{(t+1)})\right] + \frac{1}{\beta}\left((x^{(t+1)} - x^*)^T e^{(t+1)}\right) \\ &\leq \frac{1}{\beta}\left[d(x^{(t)}) - d(x^{(t+1)})\right] + \frac{1}{\beta}\left(C\|e^{(t+1)}\| + \langle x^{(t+1)}, e^{(t+1)}\rangle\right),\end{aligned}$$

where $C := \sup_{x\in X^*}\{\|x\|\}$. The second inequality is obtained through triangle inequality. Then

$$d^{(t+1)} \leq \mu d^{(t)} + \mu\left(C\|e^{(t+1)}\| + \langle x^{(t+1)}, e^{(t+1)}\rangle\right),$$

where $\mu = \frac{1}{1+\beta\eta} < 1$. With our assumptions and according to Theorem 2 and Corollary 2 in [33], we guarantee the generated sequence converges linearly in the order of $\mathcal{O}\left(c^t\right)$, where $c = \sqrt{\frac{1+\max\{\mu,\rho\}}{2}} \in (0, 1)$.

Based on the above four items, we guarantee the convergence results in this theorem. □