Dictionary Learning

We've been working with the problem:

$$\min_{\gamma} \frac{1}{2} \|y - D\gamma\|_2^2 \text{ s.t. } \|\gamma\|_0 \leq k.$$

What is the correct/best dictionary to use? We can learn it from a collection of data.

Let $Y = [y_1 \cdots y_N] \in \mathbb{R}^{n \times N}$, then:

$$\min_{\Gamma, D} \frac{1}{2} \|Y - D\Gamma\|_2^2 \text{ s.t. } \begin{cases} \|\gamma_i\|_0 \leq k \\ \|d_i\|_2 = 1 \end{cases} = \text{Dict. Learning Problem.}$$

The most common approach is <u>alternating minimization</u>:

- fix D,
$$\min_{\gamma_i} \frac{1}{2} \|y_i - D\gamma_i\|_2^2 \text{ s.t. } \|\gamma_i\|_0 \leq k.$$
 - which can be solved with OMP, relaxed to l_1, etc.

- fix Γ, solve for D:
$$\min_{D} \frac{1}{2} \|Y - D\Gamma\|_2^2 \text{ s.t. } \|d_i\|_2 = 1$$

The latter can be done in a number of different ways. Most notable, Method of Optimal Directions (MOD) finds the "optimal" D for a given Γ:

$$D = (Y\Gamma^T)(\Gamma\Gamma^T)^{-1}.$$

Alternatively, one could minimize one atom/column at a time:

$$\min_{d_j} \frac{1}{2}\| Y - D\Gamma^T \|_2^2 = \frac{1}{2} \| Y - \sum_{i \neq j} d_i \Gamma_i^T - d_j \Gamma_j^T \|$$
$$= \frac{1}{2} \| E_j - d_j \Gamma_j^T \|_2^2$$

This leads to the K-SVD ~~alg~~.

<u>Optimization Guarantees:</u>

The DL problem is non-convex, not just because of the $\|\gamma_i\|_0$, but also due to the product $D.\Gamma$. Thus, convergence to a global optimum is complicated, and usually one can guarantee convergence to a local minimum.

Some algorithms have some convergence guarantees, alas they tend to be quite involved. Thus, instead of going through one of these, we will take a simpler approach: we will revisit convergence guarantees for gradient descent, and then we will see how to use these in a non-convex case as above.

Gradient Descent

Let $f: \mathbb{R}^n \to \mathbb{R}$.

We will assume f is twice differentiable, β-smooth and α-strongly convex. Diffent guarantees can be obtained relaxing some of these.

For $t = 1$ to T,

$$x^{t+1} = x^t - \eta \nabla f(x^t).$$

end

<u>β-smoothness:</u> f is β-smooth if

$$\|\nabla f(y) - \nabla f(x)\| \leq \beta \|y - x\| \quad , \forall x, y.$$

Note that if f is twice diff, then this is equivalent to

$$\|\nabla^2 f(x)\| \leq \beta \quad \forall x.$$

<u>Strong Convexity:</u> f is α-strongly convex if

$$(y-x)^T \nabla^2 f(x) (y-x) \geq \alpha \|y - x\|^2 \quad , \forall x, y.$$

This implies that one "can fit" a quadratic function underneath $f \Rightarrow f$ is not "too flat".

Note that this also implies

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\alpha}{2} \|y-x\|^2$$

We will prove the following main result:

Thm: Let f be twice differentiable, β-smooth and α-convex. Let x^* be the (unique) minimizer of f and $\eta \leq 1/\beta$. Then Gradient Descent satisfies: (From x_1).

$$f(x_t) - f(x^*) \leq \beta \left(1 - \frac{\eta \alpha}{2}\right)^{t-1} \|x_1 - x^*\|^2.$$

In the proof, we will employ the following lemma:

Lemma: If f: twice diff, β-smooth and α-strongly convex, then (Key).

$$\boxed{\nabla f(x_t)^T (x_t - x^*) \geq \frac{\alpha}{4} \|x_t - x^*\|^2 + \frac{1}{2\beta} \|\nabla f(x_t)\|^2.}$$

"Sufficient correlation".

Proof of Thm:

For simplicity, let $\alpha' = \frac{\alpha}{4}$; $\beta' = \frac{1}{2\beta}$. Consider

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \eta \nabla f(x_t) - x^*\|^2 \\
&= \|x_t - x^*\| - 2\eta \nabla f(x_t)^T (x_t - x^*) + \eta^2 \|\nabla f(x_t)\|^2. \\
&\leq \|x_t - x^*\| - 2\eta\left(\alpha' \|x_t - x^*\|^2 + \beta' \|\nabla f(x_t)\|^2\right) + \eta^2 \|\nabla f(x_t)\|^2 \\
&= \|x_t - x^*\| (1 - 2\eta\alpha') + (\eta^2 - 2\eta\beta') \|\nabla f(x_t)\|^2
\end{aligned}$$

as long as $\cancel{\beta}\eta < 1/\beta \Rightarrow$

$$\|x_{t+1} - x^*\|^2 \leq (1 - 2\eta\alpha')\|x_t - x^*\|^2. \qquad (1)$$

Finally, note that $f(x^*) \geq f(x_t) + \nabla f(x_t)^T (x^* - x_t)$

$$\Rightarrow \quad f(x_t) - f(x^*) \leq \nabla f^T(x_t)(x_t - x^*) \leq \beta \|x_t - x^*\|^2. \qquad (2)$$

$$= \underbrace{(\nabla f(x_t) - \nabla f(x^*))^T (x_t - x^*)}_{\geq 0} \longrightarrow \beta\text{-smooth}.$$

from (1) and (2):

$$f(x_t) - f(x^*) \leq \beta(1 - 2\eta\alpha')\|x_t - x^*\|^2.$$

//.

Proof of Lemma:

Recall we need to get, under strong-convexity and smoothness assumptions

$$\nabla f^T(x_t)(x_t - x^*) \geq \frac{\alpha}{4}\|x_t - x^*\|^2 + \frac{1}{2\beta}\|\nabla f(x_t)\|^2.$$

for the first part, note that

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\alpha}{2}\|x - x^*\|^2, \quad \text{by } \alpha\text{-strong convexity}.$$

But $f(x) \geq f(x^*)$, thus: $\nabla f(x)^T (x - x^*) \geq \frac{\alpha}{2}\|x - x^*\|^2. \qquad (a)$

Then, recall the Lagrange reminder theorem:

Let $f: \mathbb{R}^n \to \mathbb{R}$, twice diff., then for some $t \in [0,1]$, $x' = ty + (1-t)x$,

$$\nabla f(x) = \nabla f(y) + \nabla^2 f(x')(x - y).$$

This can be proven by Taylor expansion of $f(x)$ and the intermediate value theorem.

Let $y = x^*$, and note that $\nabla f(x^*) = 0$. Thus,

$$\nabla f(x) = \nabla^2 f(x')(x - x^*).$$

And so:

$$\nabla f(x)^T (\nabla^2 f(x'))^{-1} \nabla f(x) = \nabla f(x)^T (x - x^*)$$

Recall that β-smoothnes implies

$$\nabla f(x)^T (\nabla^2 f(x'))^{-1} \nabla f(x) \geq \frac{1}{\beta} \|\nabla f(x)\|_2^2$$

$$\Rightarrow \nabla f(x)^T (x - x^*) \geq \frac{1}{\beta} \|\nabla f(x)\|_2^2. \qquad (b)$$

Thus, from (a), (b):

$$\nabla f(x)^T (x - x^*) \geq \frac{\alpha}{4} \|x_t - x^*\|_2^2 + \frac{1}{2\beta} \|\nabla f(x)\|_2^2$$

//.

(Consider commenting on SGD).

Following Analysis from [Arora, Ge, Ma, Moitra, Simple, Efficient neural Algorithms for S. C.].

The cool thing about this analysis is that one does not necessarily have to be taking the gradient as a direction. As long as this direction satisfies a version of the key lemma from before "sufficient correlation", the analysis follows through.

Making this more precise:

<u>Def:</u> A vector g_t is $(\alpha', \beta', \epsilon_t)$-correlated with a point x^* if, $\forall\, t$,

$$g_t^T(x - x^*) \geq \alpha' \|x_t - x^*\|^2 + \beta' \|g_t\|^2 - \epsilon_t .$$

We saw before that if f is twice diff., ~~then~~ α-strongly convex and β-smooth, then $\nabla f(x_t)$ is $(\frac{\alpha}{4}, \frac{1}{2\beta}, 0)$-correlated with the optimal solution.

Further, the proof for the previous theorem generalizes directly ~~the~~ to this case too.

<u>Thm:</u> Suppose g_t is $(\alpha', \beta', \epsilon_t)$-correlated with a point x^*, and $\eta < 2\beta'$. Then "abstract gradient descent" $x^{t+1} = x^t + \eta g^t$ from x_1 satisfies

$$\|x_t - x^*\| \leq \left(1 - \frac{\eta \alpha'}{2}\right)^{t-1} \|x_1 - x^*\|^2 + \underbrace{\frac{\max \epsilon_t}{\alpha'}}_{\text{bias}} .$$

Back To Dictionary learning.

We will assume a stochastic setting, where

a) $Supp(\gamma) \sim$ u.a.r, $\|\gamma\|_0 = k$.

b) $\gamma_{S_i} \begin{cases} +1 & wp.\ 0.5 \\ -1 & wp\ 0.5 \end{cases}$, pair-wise independent conditioned on the support S.

We observe $y = D\gamma$.

Recall that we are after

$$\min_{D,\Gamma} \|Y - D\Gamma\|_F^2 \text{ st. } \|\gamma_i\|_0 \leq k. \ \forall i.$$

Observation: instead of viewing alternating methods as minimizing a known function, we can think of them as minimizing an unknown - convex function.

In other words: the problem

$$\min_D \|Y - D\Gamma\|_2^2$$ is convex. It's just that we don't really know Γ.

What we will do is show that alternating methods still move in a direction "sufficiently correlated" with the true gradient:

$$f_1(D) = \|Y - D\Gamma\|^2 \approx f_2 = \|Y - D\hat{\Gamma}\|^2$$

and we expect $\nabla f_2(D) \approx \nabla f_1(D)$

have know — we don't know.

~~It moves almost in the ideal direction of the true parameter~~ [illegible] (5)(6)

We will analyze the simple algorithm:

for $t=0$ to T:

- "Decoding": $\hat{\gamma}^{(i)} = \mathcal{H}_{1/2}(\hat{D}^T y^{(i)}) \quad \forall i$.
- "Update": $\hat{D} \leftarrow \hat{D} + \eta \sum_{i=1}^{N} (y_i - \hat{D}\hat{\gamma}_i)\, \mathrm{sign}(\hat{\gamma}_i)^T$.

<u>Metric / Distance</u>:

Two matrices $D, \hat{D}$, (with normalized columns) are (δ, κ)-close if $\exists$ permutation & sign flip of the columns of $\hat{D}$: $B = \hat{D}P$ such that $\|b_i - d_i\|_2 \leq \delta \quad \forall i$

and $\|B - D\| \leq \kappa \|D\|$.

<u>Decoding Succeeds</u>:

Assume $D: n \times m$ is μ-incoherent and $b = D\gamma$, with $\|\gamma\|_0 = k \leq \frac{1}{10\mu \cdot \log(n)}$, and $\hat{D}$ is $\left(\frac{1}{\log n}, 2\right)$-close to D.

Then, the decoding stage succeeds: (with high probability).

$$\mathrm{sign}\left[H_{1/2}(\hat{D}^T b)\right] = \mathrm{sign}(\gamma).$$

A proof sketch is the following:

Consider the j^{th} inner-product:

$$\langle \hat{d}_j, b\rangle = \hat{d}_j^T D\gamma = (\hat{d}_j - d_j + d_j)^T d_j \gamma_j + \hat{d}_j^T \sum_{j\neq i} d_i \gamma_i$$

$$= \gamma_j + \underbrace{(\hat{d}_j - d_j)^T d_j \gamma_j}_{\|\cdot\| \leq \frac{1}{\log(n)}} + \underbrace{\sum_{j\neq i} \hat{d}_j^T d_j \gamma_i}_{\ll}$$

So w.h.p., if $j \in S$, $|\hat{d}_j^T b| > 1/2$, and $|\hat{d}_j^T b| < 0$ otherwise.

<u>Update:</u>

We'll assume the update is given by

$$g = \mathbb{E}\left((y - \hat{D}\hat{\gamma})\, \mathrm{sign}(\hat{\gamma})^T\right),$$

and the j^{th} atom is updating through $g_j = \mathbb{E}[(y - \hat{D}\hat{\gamma})\, \mathrm{sign}(\hat{\gamma}_j)]$.

Denote by F the event that encoder recovers the support of γ, which holds w.h.p. if $\hat{D}, D$ are $(1/\log n, 2)$-close. Then:

$$g_j = \mathbb{E}[(y - \hat{D}\hat{\gamma})\, sg(\hat{\gamma}_j) \mathbb{1}_F] + \mathbb{E}[(y - \hat{D}\hat{\gamma})\, sg(\hat{\gamma}_j) \mathbb{1}_{\bar{F}}]$$

$$= \mathbb{E}[(y - \hat{D}\hat{\gamma})\, sg(\hat{\gamma}_j) \mathbb{1}_F] \pm \epsilon \quad , \text{ say } \|\epsilon\| < n^{-c}.$$

so $$g_j = \mathbb{E}[(y - \hat{D} H_0(\hat{D}^T y))\, sg(\gamma_j) \mathbb{1}_F] \pm \epsilon \quad ; \text{ say } S = supp(\gamma)$$

$$= \mathbb{E}[(I - \hat{D}_S \hat{D}_S^T) y \cdot sg(\gamma_j) \mathbb{1}_F] \pm \epsilon$$

$$= \mathbb{E}[(I - \hat{D}_S \hat{D}_S^T) D_S \gamma \cdot sg(\gamma_j)] \pm \epsilon.$$

ow, using subconditioning: (First sampling the support S and then the values γ_S),

$$g_j = \mathbb{E}_S\Big[\mathbb{E}_{\gamma_S}\big[(I - \hat{D}_S D_S^T) D\gamma \cdot \mathrm{sgn}(\gamma_j) \mid S\big]\Big] \pm \epsilon$$

$$= \mathbb{E}_S\Big[\mathbb{E}_{\gamma_S}\big[(I - \hat{D}_S \hat{D}_S^T) D_j \gamma_j\, \mathrm{sgn}(\gamma_j) \mid S\big]\Big] \pm \epsilon \quad \text{because } \gamma_{S_i} \text{ are iid given } S. \text{ or uncorrelated.}$$

$$= \mathbb{E}\big[p_j \cdot (I - \hat{D}_S \hat{D}_S^T)\, d_j\big] \pm \epsilon, \quad p_j = \mathbb{E}[\gamma_j \cdot \mathrm{sgn}(\gamma_j) \mid S].$$

$$= p_j \cdot \mathbb{E}\big[(I - \hat{D}_S \hat{D}_S^T)\, d_j\big]. \pm \epsilon$$

$$= p_j\, \mathbb{E}\big[(I - \hat{d}_j \hat{d}_j^T)\, d_j\big] + p_j\, \mathbb{E}\big[\hat{D}_{-j} \hat{D}_{-j}^T d_j\big] \pm \epsilon.$$

$$= \underbrace{p_j q_j}_{q_j = P[j \in S]} (I - \hat{d}_j \hat{d}_j^T)\, d_j + p_j \hat{D}_{-j} \underbrace{Q}_{\mathrm{diag}(\{q_{i,j}\}_i)} \hat{D}_j^T d_j \pm \epsilon.$$

Note that then
$$g_j = p_j q_j \big(d_j - \underbrace{\lambda_j}_{\hat{d}_j^T d_j \simeq 1} \hat{d}_j\big) + \underbrace{\varphi_j}_{\|\varphi_j\| \le O(k/n)} \pm \epsilon.$$

This "shows" that g_j is mostly correlated with the optimal direction $(d_j - \hat{d}_j)$:

__Lemma__: If $g_j = 4\alpha(\hat{d}_j - d_j) + v$, with $\|v\| \le \alpha \|\hat{d}_j - d_j\| + \xi$, then g_j is $(\alpha, 1/100\cdot\alpha, \xi^2/\alpha)$-correlated with d_j^*

$\Rightarrow$ g_j is $(\alpha, \beta, \epsilon)$-correlated with d_j, with $\alpha = \Omega(k/m)$, $\beta \ge \Omega(m/k)$, $\epsilon = O(k^3/mn)$. $\Rightarrow$ Each Grad. Step makes progress.

Corollary:

$O(1/\log n)$

If $\hat{D}$ is $(2\delta, 2)$-near D, and $\eta \leq \min_i (p_i q_i (1-\delta))$, then

$$\|\hat{d}_j^{k+1} - d_j\|^2 \leq (1 - 2\alpha\eta)\, \| d_j^k - d_j \|^2 + O(k/n).$$

$\Rightarrow$ $\hat{D}_j$ converges geometrically to d_j until the column-wise error is $O(\sqrt{k/n})$.