Mirror descent over measures

Let $\mathcal{X} \subset \mathbb{R}^d, \mathcal{M}(\mathcal{X})$ the space of Radon measures on $\mathcal{X}$, convex functionals $F, \phi : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$, convex $C \subset \mathcal{M}(\mathcal{X})$, consider mirror descent:

$$\min_{\mu \in C} F(\mu)$$

Equation (1)

Under which assumptions does it converge and for which rate?

Examples of optimization of measures

The "Kullback-Leibler divergence" or relative entropy is

$$KL(\mu | \nu) = \int \log \left(\frac{\mu(x)}{\nu(x)}\right) d\nu(x)$$

Entropic optimal transport $\min_{\nu \in C, \mu \in \mathcal{M}(\mathcal{X})} KL(\nu | \mu)$ for $P \propto \exp(-\langle \xi, x \rangle) | \mu \propto \mathcal{N}(0,1)$

Expectation-Maximization $\min_{\nu \in C} KL(\nu | \mathcal{P}(\bar{X}))$ with the observations $\bar{X}$

Bayesian inference $\min_{\nu \in C} KL(\nu | \mu)$ with the posterior $\mu \propto \exp(-V)$

Optimization of 1-hidden layer neural network $\min_{\nu \in C} \text{MMD}(\nu | \mu)$

Definitions of derivatives and relative smoothness

The KL does not have a Gâteaux derivative! Need weaker noces:

$$(\text{directional derivative}) d^* F(\mu)(\nu - \mu) = \lim_{h \to 0} \frac{F(\nu + h\nu) - F(\nu)}{h}$$

$$\langle \nabla F(\mu), \xi \rangle = d^* F(\mu)(\xi)$$

$$D_{\nu}(\mu) = (\phi(\nu) - \phi(\mu)) - d^* F(\mu)(\nu - \mu)$$

Mirror descent over measures

Convergence result for mirror descent

Theorem: Assume that $F$ is $l$-strongly convex and $L$-smooth relative to $\phi$, with $l, L \geq 0$. Consider the mirror descent scheme (1), and assume that for each $n \geq 0$, $\nabla F(\nu_n)$ exists. Then for all $n \geq 0$ and all $\nu \in \text{dom}(F) \cap \text{dom}(\phi)$

$$F(\mu_n) - F(\nu) \leq \frac{1}{2} D_{\nu}(\mu_n) + \left(\frac{1}{1 + e^{L/\phi}}\right) - 1 \leq \frac{L}{n} D_{\nu}(\mu_0)$$

Entropic optimal transport and Sinkhorn

For any $n \in \mathbb{N}$, we can write $\pi = p_X \otimes K_n$, where $K_n(x,dy) \propto e^{\phi(x) - \phi(y)}$. Hence we have the decomposition:

$$KL(\pi | \bar{\pi}) = \int \log \left(\frac{\pi(x)}{\bar{\pi}(x)}\right) d\bar{\pi}(x) = KL(\pi | \bar{\pi}) + KL(\bar{\pi} | \pi)$$

Define cyclically invariant $\pi_n \in C$, if for $(\mu, \nu) = (p_X, \pi_n)$ its marginals,

$$KL(\pi | e^{\phi/\mu} \pi) = \min_{n \in \text{dom}(\phi)} KL(\pi | e^{\phi/\nu} \pi)$$

When $\pi_n$ is unique, there exist $f, g \in L^\infty(\mathcal{X} \times \mathcal{X})$ such that $e = e^{fg/\mu} \pi$$

Main result for EM

Proposition: For all $n \geq 0$, the EM algorithm is a mirror descent and verifies, for $\pi_n$ the optimum of EM and $\mu_n$, its first marginal,

$$KL(\mu_n | \pi_n) \leq \frac{KL(\pi_0 | \pi_n)}{(1 + e^{L/\phi}) - 1}$$

EM and latent EM

We posit a joint distribution $p_{xy}(dx,dy)$ parametrized by an element $q$ of some given set $Q$. For $p_{xy}(dx,dy) = \int p_{xq}(dx,dq)$, the goal is to infer $q$ by solving

$$\min_{q \in Q} KL(p_{xy} | \nu_q)$$

Main result for general EM

Proposition: EM is a mirror descent, with $\nabla F_{\text{EM}}(\pi; q) = \ln(d\pi_n/d\nu_q)$

Proof: Use the envelope theorem to differentiate $F_{\text{EM}}$ and find that $\nabla F_{\text{EM}}(\pi,q) = \ln(d\nu_q/d\nu_n)$. Then for any coupling $\pi$, we have the identity

$$F_{\text{EM}}(\pi) = \inf_{q \in Q} KL(\pi | \nu_q)$$

Main result for latent EM

Proposition: Latent EM can be written as mirror descent with objective $F_{\text{EM}}$: Bregman potential $\phi$ and the constraints $C = \{\pi_0, \nu\}$

$$\pi_{n+1} = \arg\min_{\nu \in \text{dom}(\phi)} F_{\text{EM}}(\pi_n, \nu) = \inf_{\nu \in \text{dom}(\phi)} KL(\nu | \nu_0)$$

\[\text{Lemma:}\] The function $F_{\text{EM}}$ is convex and $1$-relatively smooth w.r.t. the negative entropy $\phi$, over $\mathcal{P}(\mathcal{X})$.

\[\text{Consequence:}\] This already yields a $O(1/n)$ rate for Sinkhorn's algorithm.

Proposition: Set $\mu_n \in \arg\min_{\nu \in \text{dom}(\phi)} KL(\nu | \nu_0)$. Then $\mu_n$ is $1$-relatively smooth to $\phi$, i.e.,

$$F_{\text{EM}}(\pi_{n+1}) \leq 1 + 4L/\phi$$

\[\text{Consequence:}\] This yields a linear rate for Sinkhorn's algorithm.