1 Math notes

Taylor series multivariate.

$$f(\boldsymbol{r}_0 + \boldsymbol{a}t) = f(\boldsymbol{r}_0) + [\boldsymbol{a} \cdot \nabla f(\boldsymbol{r})]\Big|_{\boldsymbol{r}=\boldsymbol{r}_0} t + \frac{1}{2!}[\boldsymbol{a} \cdot \nabla][\boldsymbol{a} \cdot \nabla] f(\boldsymbol{r})\Big|_{\boldsymbol{r}=\boldsymbol{r}_0} t^2 + \dots \tag{1}$$

Radon Nikodym derivative Derivative between probability distributions.

Shannon Entropy.

Self information of event x = x is defined as $I(x) := -\log P(x)$

$$H(x) = \mathbb{E}_{x\sim P}\left[I(x)\right] = -\mathbb{E}_{x\sim P}\left[\log P(x)\right]$$

Cramer-Rao lower bound. [Balabdaoui and van de Geer, 2016] Suppose θ is an unknown deterministic parameter which is to be estimated from measurements x, distributed according to some pdf $f(x;\theta)$. The variance of any *unbiased estimator* $\hat{\theta}$ of θ is then bounded by reciprocal of Fischer Information $I(\theta)$:

$$\begin{aligned}\text{var}(\hat{\theta}) &\geq \frac{1}{I(\theta)} \text{ where} \\ I(\theta) &= \mathbb{E}\left[\left(\frac{\partial l(x;\theta)}{\partial \theta}\right)^2\right] \\ &= -\mathbb{E}\left[\frac{\partial^2 l(x;\theta)}{\partial \theta^2}\right]\end{aligned}$$

Note: See Wikipedia for other more general versions

Lipschitz continuity, smoothness, etc.. f is $M-$*Lipschitz continuous* given M if

$$|f(x) - f(y)| \leq M|x - y| \forall x, y \in \mathbb{R}$$

. If f is differentiable, Lipschitz continuity says that f has bounded derivative.

f is L-*Lipschitz smooth* if its derivatives are Lipschitz continuous with L. This is called smoothness type $C^{1,1}$ i.e

$$\forall x, y \in \mathbb{R}, ||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$$

The definition does not assume convexity of f. Some other equivalent conditions are: on here.

$$g(x) = \frac{L}{2}x^\intercal x - f(x) \text{ is convex} \tag{2}$$

$$f(y) \leq f(x) + \nabla f(x)^\intercal (y - x) + \frac{L}{2}||y - x||^2, \forall x, y \tag{3}$$

$$(\nabla f(x) - \nabla f(y))^\intercal (x - y) \leq L||x - y||^2, \forall x, y \tag{4}$$

$$f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y) - \frac{\alpha(1-\alpha)L}{2}||x - y||^2 \forall x, y \in \mathbb{R}, \alpha \in [0, 1] \tag{5}$$

$$\dots \tag{6}$$

Strong Convextiy: f is α-strongly convex if

$$\nabla^2 f(x) \succeq \alpha I \forall x$$

Also

$$f(x + y) \geq f(x) + y^\intercal \nabla f(x) + \frac{\alpha}{2}||x - y||^2 \tag{7}$$

7 is equivalent to saying $g(x) = f(x) - \frac{\alpha}{2}||x||^2$ is convex.. Latter is equivalent to $\nabla^2 g \succeq 0 \equiv \nabla^2 f \succeq \alpha I$. more info here

Note:l- strong convexity can be written as

$$f(y) \geq f(x) + \nabla f(x)^\intercal (y - x) + \frac{l}{2}||y - x||^2 \tag{8}$$

Equation 8 is direct contrast to 3. $\frac{L}{l}$ is called condition number of matrix.

Functional Optimization on KL divergence.

in progress

In function space, dot product is extended as $\langle \cdot, \cdot \rangle$

$$\langle f, g \rangle = \int f(\theta) g(\theta) d\theta$$

Now,

$$\begin{aligned} \text{KL}\left(q||p\right) &= \mathbb{E}_q\left[\log q\right] - \mathbb{E}_q\left[\log p\right] \\ &= \langle q, \log \frac{q}{p} \rangle \\ &=: f(q) \end{aligned} \tag{9}$$

For functional f, derivative defined as

$$f(q + \epsilon d) = f(q) + f'(q)\epsilon + O(\epsilon^2) \tag{10}$$

$$f'(q) = \int \frac{\partial f(q)}{\partial q(\theta)} d(\theta) d\theta \tag{11}$$

$$\equiv \langle \nabla f(q), d \rangle \tag{12}$$

From equation 9,

$$\nabla f(q) := \frac{\partial f(q)}{\partial q} = \log \frac{q}{p} \tag{13}$$

In finite dimensional optimization, we look for candidates for optima by looking for $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0$. This is equivalent to asking every component of $\nabla f \stackrel{!}{=} 0$. For functions we look for q^* s.t

$$\frac{\partial f(q^*)}{\partial q(\theta)} \stackrel{!}{=} 0 \forall x \in \mathbb{R}$$

Frank Wolfe for Variational Inference Optimization.

in progress

Algorithm **??** where optimization is over probability distributions. q_t, s in [Locatello et al., 2018] is $\equiv$ $\mathbf{x}_t, \mathbf{s}$ in all discussion on Frank-Wolfe. One is a function while the other is a vector. $\mathbf{x}_t$ can also be seen as parameters of the probability distributions but we do not use the same distance functions and gradients Objective function so it would be better to stay in function space.

Objective function f is $f(\mathbf{x}_t) \equiv f(q_t) = \mathcal{D}^{kl} q_t || p$ where p is the target distribution. $\mathbf{d}_t = \mathbf{s}_t - \mathbf{x}_t$. $||\mathbf{d}_t||$ can be replaced by a divergence metric between s and q_t like $\text{KL}\left(s_t||q_t\right)$. We can also try $\text{KL}\left(q_t||s\right)$ or others metrics. Gap g_t is always scalar.

is there an analog for this in functional space?

$$\begin{aligned} g_t &= -\langle \nabla f(\mathbf{x}_t), \mathbf{d}_t \rangle \geq 0 \\ &= \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t \rangle - \langle \nabla f(\mathbf{x}_t), \mathbf{s}_t \rangle \end{aligned} \tag{14}$$

For probability distribution q,

$$\langle h, q \rangle = \langle q, h \rangle = \mathbb{E}_q\left[h\right] \tag{15}$$

$$\approx \frac{1}{S} \sum_{\theta_i \sim q}^{i=1..S} f(\theta_i) \tag{16}$$

i.e expectation of f under q. 14 then becomes

$$g_t = \mathbb{E}_{q_t}\left[\nabla f(q_t)\right] - \mathbb{E}_{s_t}\left[\nabla f(q_t)\right] \tag{17}$$

$$= \mathbb{E}_{q_t}\left[\log \frac{q_t}{p}\right] - \mathbb{E}_{s_t}\left[\log \frac{q_t}{p}\right] \tag{18}$$

2 Code Notes

Normal distribution edward

```
from edward.models import Normal
from keras.layers import Dense

hidden = Dense(256, activation='relu')(x_ph)
qz = Normal(loc=Dense(10)(hidden),
scale=Dense(10, activation='softplus')(hidden))
```

`Edward` has issues with `Tensorflow versions > 1.7` see #893. Issue is non-trivial and it doesn't seem like there is a plan to fix since `Edward2` is part of `tensorflow_probability`.

3 Web pages

1. Zen of gradient descent with intro to Nexterov Method

2. I am a bandit Nesterov accelerated

3. CMU Stats FW lecture

References

[Balabdaoui and van de Geer, 2016] Balabdaoui, F. and van de Geer, S. (2016). Fundamentals of mathematical statistics.

[Locatello et al., 2018] Locatello, F., Dresdner, G., Khanna, R., Valera, I., and Rätsch, G. (2018). Boosting black box variational inference. *arXiv preprint arXiv:1806.02185*.