Mirror Descent Methods (MD)

问世于1979年[1]，打造这一类型算法的目的即为求解高维问题，这从[1]的标题中可见一斑。

在相同作者于1983年出版的书[2]，有下述论述：

"A special feature of these methods is that their laboriousness does not depend explicitly on the dimension of the problem. Accordingly, it is sensible to use them for solving convex problems of high dimension."

"Number of steps in the work of certain ~~work~~ meth on certain problem."

"For economics problems, the requirements regarding the accuracy of solution are usually not too great, whereas their dimensionality may be very considerabl

Therefore, the construction of convex-programming methods which are "not sensitive to dimensionality' is quite a pressing question."

"Clearly, it is impossible to put forward a method of solving general convex problems with a bound for the laboriousness independent of the dimension, ~~[illegible]~~ unless definite hypotheses are made about the affine properties of 可行域 X."

"Thus, in constructing a method of convex opt* insensitive to dimensionality, we have somehow or other to distinguish the necessary affine properties of $\mathcal{X}$".

[3]中MD的形式已非MD最初的样子
新的形式是更加的简洁(包括证明过程)
但最初的形式更能体现MD的核心理念
所以我先勾勒MD最初的样子，与我的理解
不一定对，大家要批判地看.

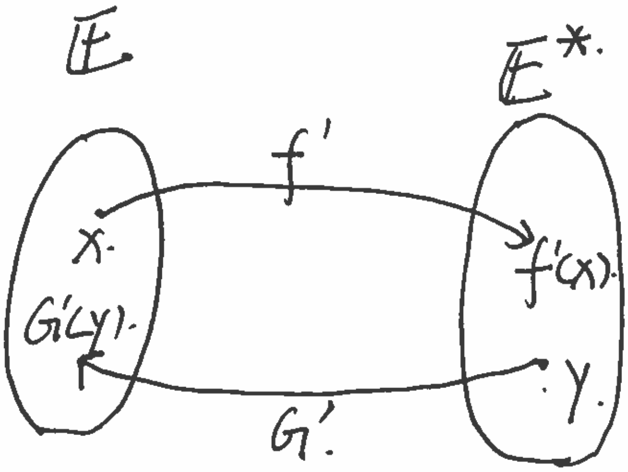
$$\min_{x \in E} f(x).$$

E是一个Banach Space，作为一个vector space
其dual space E^*是由所有E上面的
Linear Functional所组成.

MD的思想就是我们不在E中找x^*(最优
了，我们在E^*中找一个序列，或者说路径.
路径的终点，映射回E，就是x^*.

因为这样可以规避掉一些来自 $\mathbb{E}$ 的 dimension-related 的束缚.

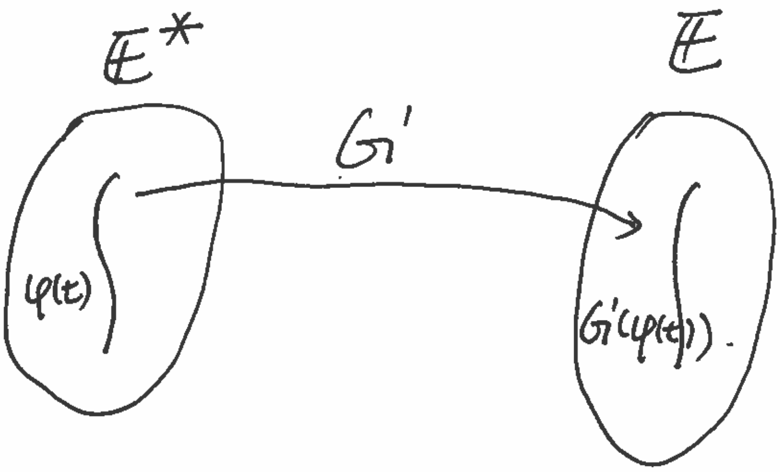
更具体地看.



$f: \mathbb{E} \to \mathbb{R}$ $f'(x)$ is a linear functional on $\mathbb{E}$, 所以 $f'(x) \in \mathbb{E}^*$.

$G: \mathbb{E}^* \to \mathbb{R}$, $G'(y)$ is a linear functional on $\mathbb{E}^*$, 所以 $G'(y) \in (\mathbb{E}^*)^* = \mathbb{E}$.

(假设 $\mathbb{E}$ is reflectiv



考虑 E^* 上面的一条路径 $\varphi(t)$, such th

$$\frac{d\varphi}{dt} = -f'(G'(\varphi(t)))$$

Define $G_*(\varphi) = G(\varphi) - \langle \varphi \mid x^* \rangle$

有 $\dfrac{d\,G_*(\varphi(t))}{dt} = \cdots \leq f(x^*) - f(x(t)) \leq$

这在说 G_* 沿着 $\varphi(t)$ 递减.

因为在选择 G 时, G 满足一定性质.

所以 $G_*(\varphi(t))$ 不会一直降落.

所以 $\overline{\lim\limits_{t\to\infty}} \dfrac{d\,G_*(\varphi(t))}{dt} = 0$.

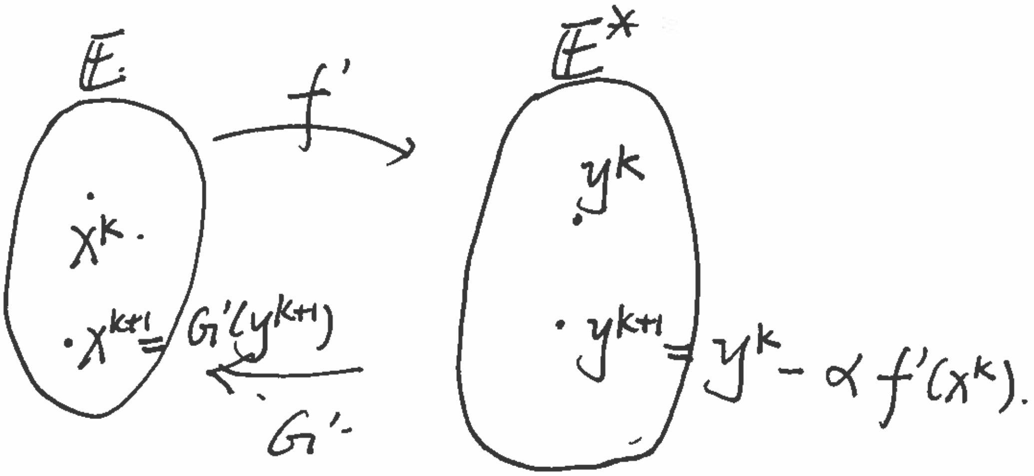
加上 f 的 Lipschi

所以 $\lim\limits_{t\to+\infty} f(x(t)) = f(x^*)$. (逼着 $x(t)$ 靠近

下一步就是将 $\frac{d\varphi}{dt} = -f'(G'(\varphi(t)))$

离散化，

也就是说，设计 $\{\varphi^k\}$，使得 $\{\varphi^k\}$ 近似 $\varphi(t)$



上面其实呈现了 MD 的主要迭代.

值得注意的是，迭代步数（也就是一开始关心的对象），

决定于

f 的 Lipsthitz constant.

accuracy 程度 ν

$\|x^*\|$

"the modulus of continuity of G' in a suitable ball",

"the rate of growth of G at infinity".

上述五项中末两项与 dimension of $\mathbb{E}$ 毫无关联。(因 G 是定义于 $\mathbb{E}^*$ 的函数)，由此奠定 ~~NEG~~ MD 算法对于 problem dimens 的相对不敏感性。

MD 与 Gradient ~~Desecnt~~ Descent 的关系

当 $\mathbb{E}$ 为一个 Hibert Space 时，$\mathbb{E}^*$ 在同构的意义下与 $\mathbb{E}$ 等价。并将 G 函数定义为 $\frac{1}{2}\|\cdot\|_*^2$，$\|\cdot\|_*$ 为 $\mathbb{E}$ 中范数 $\|\cdot\|$ 的 dual norm。

那么此时的 MD 即为 GD。

换言之，GD 是 MD 中的一个特例。那为什么不干脆用 GD 呢？因为 MD 中比如函数 G，范数的其它选择，会产生比 GD 在维度依赖方面更好的表现。

所以，在[3].[4]中，将MD与GD进行比较.（MD: Mirror descent; GD: gradient descent）的实质是将选择权更大的（指 Mapping G. $\|\cdot\|$ 等）MD与特殊形势下的MD即GD之间的比较.

MD的新形式：在欧式空间 $\mathbb{R}^n$ 中

我们知道 GD: $x^{k+1} = x^k - \eta \nabla f(x^k)$.

等价于 $$x^{k+1} = \arg\min_{x} \underbrace{f(x^k) + \langle \nabla f(x^k), x - x^k \rangle}_{f(x)\text{在}x^k\text{处的一阶泰勒展}} + \frac{1}{2\eta}\|x$$

当 $\eta = \frac{1}{L}$ 时（L为 Lipschitz constant）

为 $f(x)$ 的一个二阶多项式 Quadratic Upper bound.

1) 当第三项 $\frac{1}{2\eta}\|x - x^k\|_2^2$ 的范数变为 $\frac{1}{2\eta}(x-x^k)\nabla^2 f(x^k)(x-x^k)$. 则变为 Newton's Method.

② 当第三项变为人为定义的 $V(x,z)$ 距离函数，则为 MD. 如[3].[4].

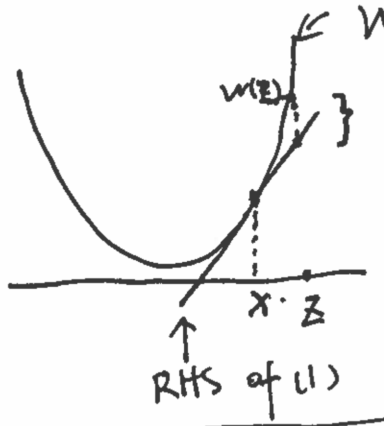
特别地，[3]中.

$$V(x,z) = w(z) - \left[w(x) + \nabla w(x)^T (z-x)\right] \quad (1)$$

所选定 $w(x)$ 被称为 distance-generating function.

it is strongly convex with parameter α, w.r.t $\|$

这样，$V(x,z)$ 可形象地认为是



这个间距即被定义为 x 与 z 之间的距离.

---

[3]中即将 MD 的这一在 $\mathbb{R}^n$ 上的新形式加以拓展至 Stochastic Setting.

除此之外，我猜测 SA 与 SAA 在 solution 精度类似而所需 Sample size (= number of SA steps) 上一直处于下风，因此[3]用较有潜力的、

较有胜算的 MD，代表 SA，出来与 SAA 进行比较。（理论层面见 [3]，1587页，以及 Numerical section）.

而此处的对比是 [3] 的另一核心着眼点。

“一直处于下风”一语并不妥当。其实 [3] 开头说明 SAA 所擅长的是 Two-Stage Sto. opt. problem 所以比较所用的例子为 minmax, saddle point problem. 之类避开胜之不武之嫌。

Den

2021. 10.

References.

[1] Nemirovsky, A.S., and Yudin, D.B. (1979)

Efficient methods of solving convex programming problems of high dimensionality.

[2] —

problem complexity and method efficiency in Optimization.

[3] —

Robust Stochastic approx. approach to SO.

[4] ~~⊗~~. Beck A. and Teboulle M.

Mirror descent and nonlinear projected subgradient methods for convex opt.